

Cycle

« Analyse de données de séquençage à haut-débit »

Module 1: Analyses ADN

9 et 10 Mars 2026

Hélène Touzet (Bonsai - CRIStAL)

Isabelle Guigon (Bilille - PLBS)

Ségolène Caboche (Bilille - PLBS)

helene.touzet@univ-lille.fr, isabelle.guigon@univ-lille.fr, segolene.caboche@univ-lille.fr

Jour 1 (13h30 -17h)

Après-midi (13h30-17h00)

- Cours
 - NGS Introduction
 - Reads Quality Control + Cleaning
- TP FastQC + multiqc + cleaning

Jour 2 (9h - 17h)

Matin (9h-12h30)

- Cours
 - Reads mapping on reference
- TP Mapping
 - Deep dive into Bowtie2 alignment parameters
- TP mapping
 - Study of a plasmid carrying antibioresistance genes

Après-midi (13h30-17h)

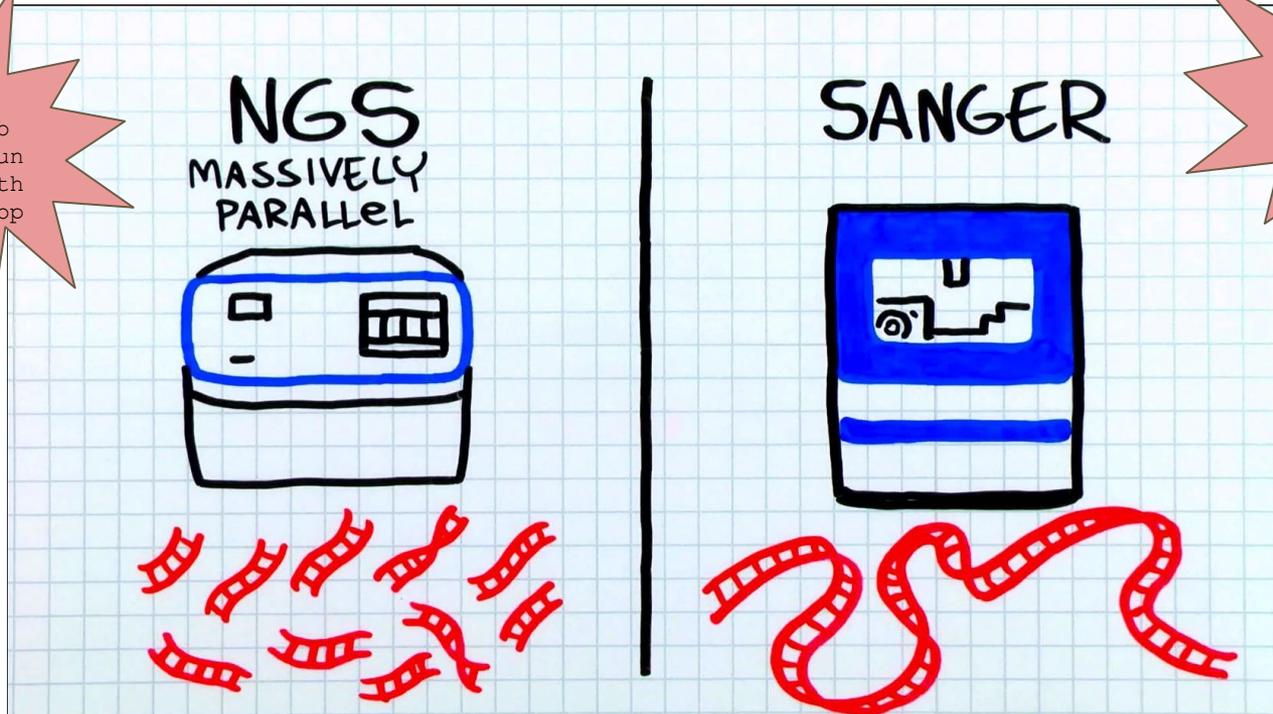
- Cours
 - Genome assembly
- TP Genome assembly
- Genome annotation

NGS introduction

What is Next-Generation Sequencing (NGS)?

“Next-generation sequencing (NGS), also known as high-throughput sequencing, is the catch-all term used to describe a number of different modern sequencing technologies. These technologies allow for sequencing of DNA and RNA much more **quickly and cheaply** than the previously used Sanger sequencing, and as such revolutionised the study of genomics and molecular biology”

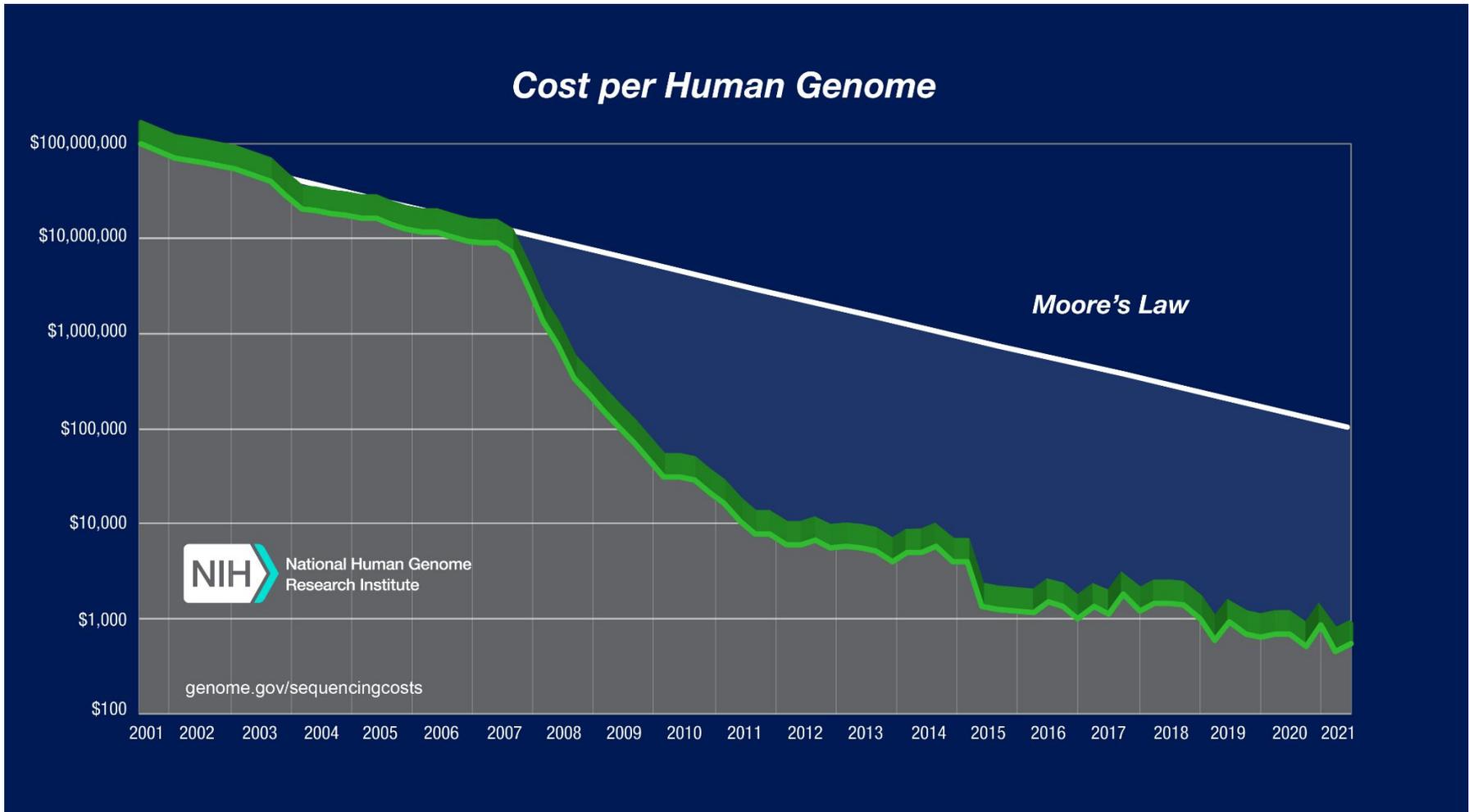
x **Gb**/run
~ 0.4 \$/Mb
~ 3 Day/run
read length
= 50~300 bp



x **Kb**/run
~ 5000 \$/Mb
~ 1 Day/run
read length
= 700 bp

The Human Genome Project was a **13-year-long** & cost **\$5 billion**

What is Next-Generation Sequencing (NGS)?



The first generation of sequencing

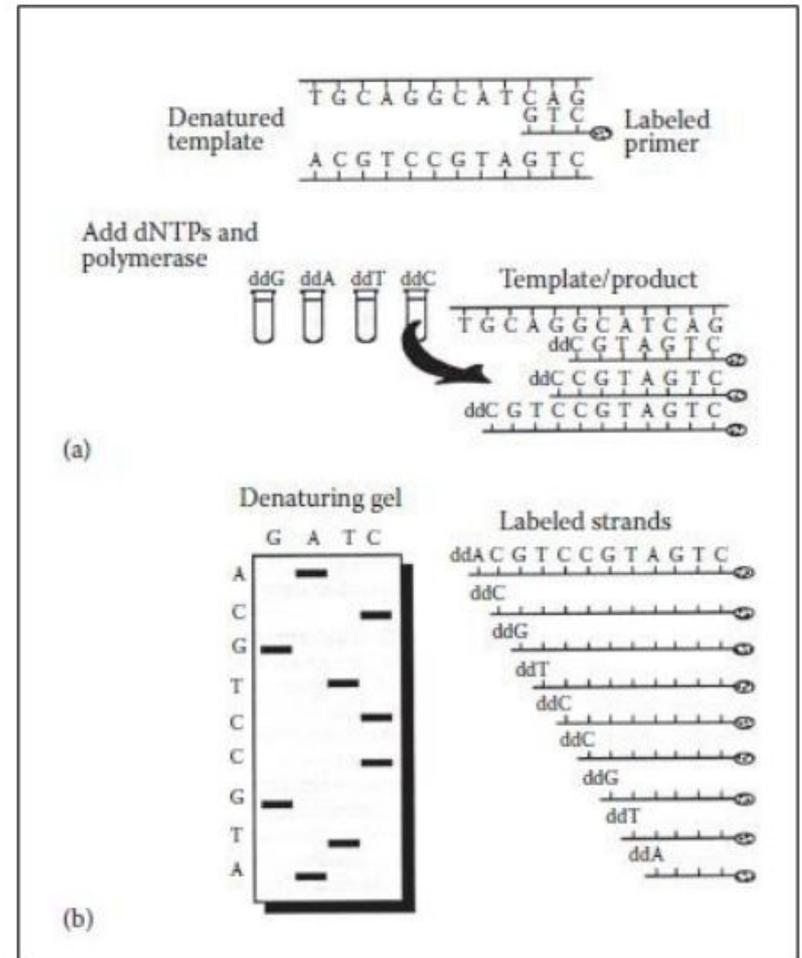
Sanger sequencing

It consists in using one strand of the double stranded DNA as template to be sequenced.

This sequencing is made using chemically modified nucleotides called **dideoxy-nucleotides** = dNTPs (ddG, ddA, ddT, and ddC).

Once incorporated into the DNA strand they prevent the further elongation and the elongation is complete => **DNA fragments ended by a dNTP with different sizes.**

The fragments are separated according to their size using gel slab where the resultant bands corresponding to DNA fragments can be visualized by an imaging system (X-ray or UV light).



What is Next-Generation Sequencing (NGS)?

The second generation of sequencing

In 2005 and in subsequent years, have marked the emergence of a new generation of sequencers to break the limitations of the first generation. The basic characteristics of second generation sequencing technology are:

- (1) The generation of many **millions of short reads in parallel**
- (2) The **speed up** of sequencing the process compared to the first generation
- (3) The **low cost** of sequencing
- (4) The sequencing output is directly detected **without the need for electrophoresis.**

The second generation of sequencing

Short read sequencing approaches divided under two wide approaches:

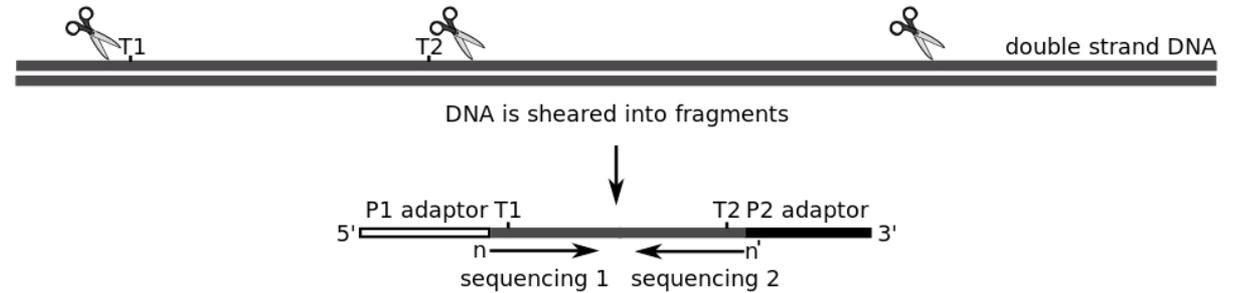
- sequencing by ligation (SBL) ABI/SOLID
- sequencing by synthesis (SBS)

and are mainly classified into four major sequencing platforms:

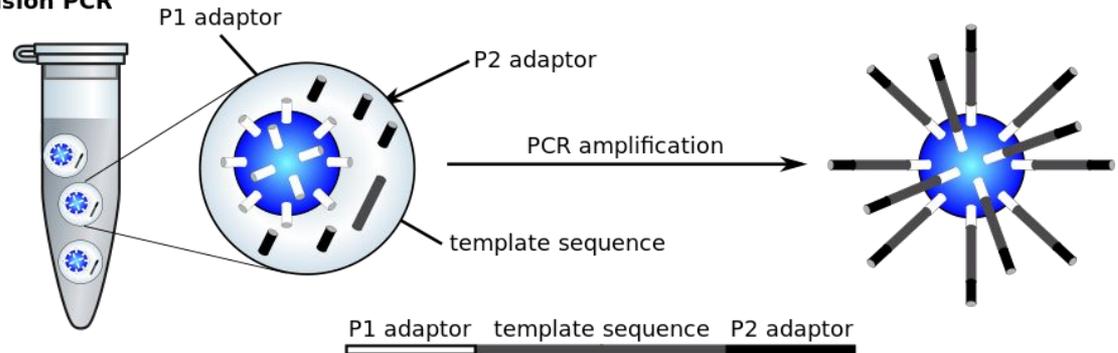
- Roche/454 launched in 2005
- Illumina/Solexa in 2006
- ABI/SOLiD in 2007
- Ion Torrent/Thermo Fisher in 2010

ABI/SOLID

(A) Single-end and paired-end sequencing



(B) Emulsion PCR



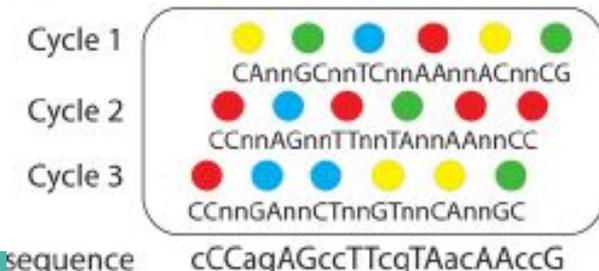
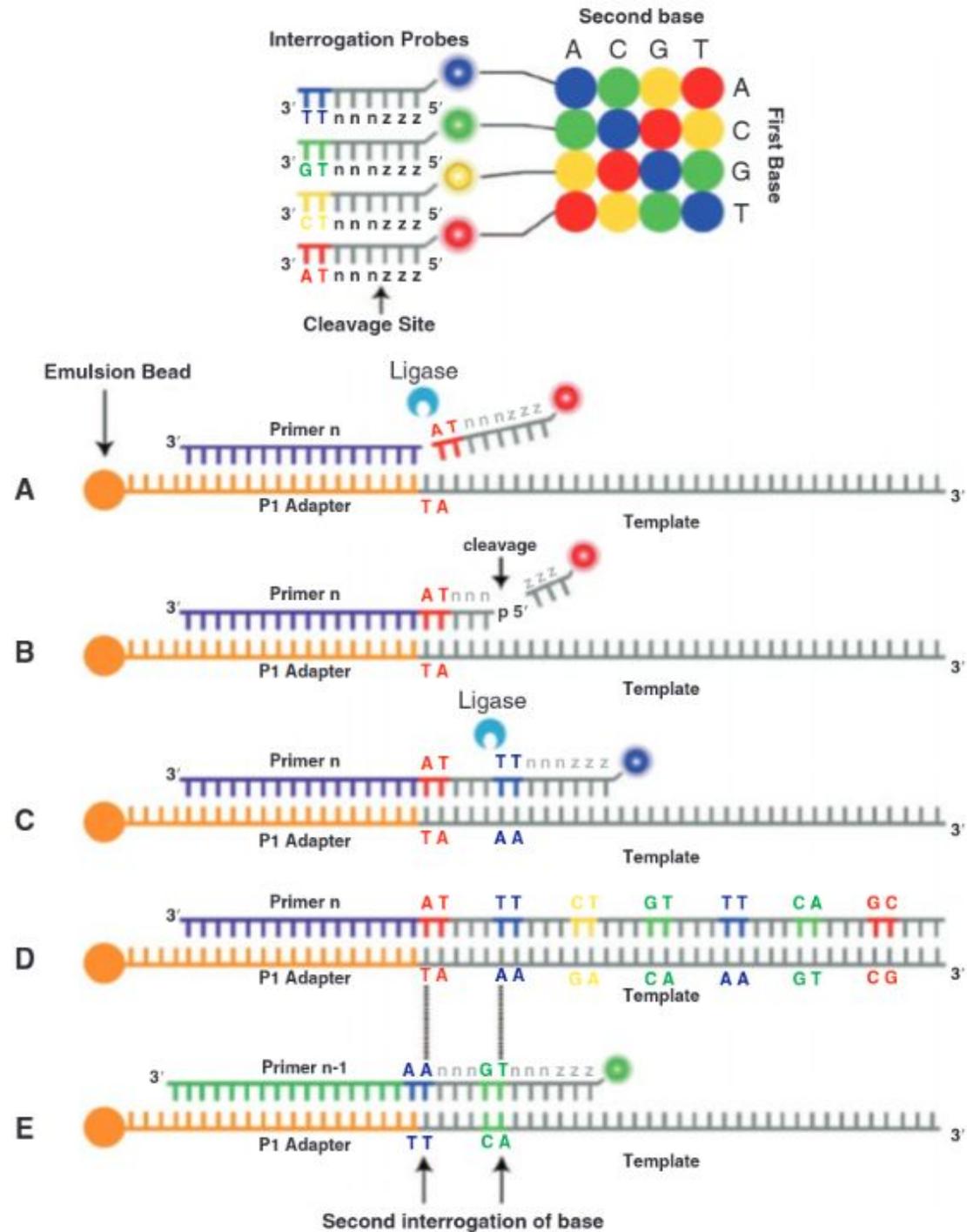
A library of DNA fragments is prepared from the sample to be sequenced, and is used to prepare clonal bead populations. That is, only one species of fragment will be present on the surface of each magnetic bead. **The fragments attached to the magnetic beads will have a universal P1 adapter sequence attached so that the starting sequence of every fragment is both known and identical.** Emulsion [PCR](#) takes place in microreactors containing all the necessary reagents for PCR. The beads with the resulting PCR products are deposited to a glass slide.

ABI/SOLID sequencing platform has been discontinued since 2016

ABI/SOLID

A set of four fluorescently labeled di-base probes compete for ligation to the sequencing primer. Interrogating every first and second base in each ligation reaction will specify the di-base probe. **Multiple cycles of ligation, detection, and cleavage are performed** with the number of cycles determining the eventual read length. After multiple rounds of ligation cycles, the extended product is removed, and this template is again set with primers corresponding to n-1 position

Available from:
archgate.net/publication/335867452 Introduction to Nucleic Acid Sequencing



Roche/454 sequencing

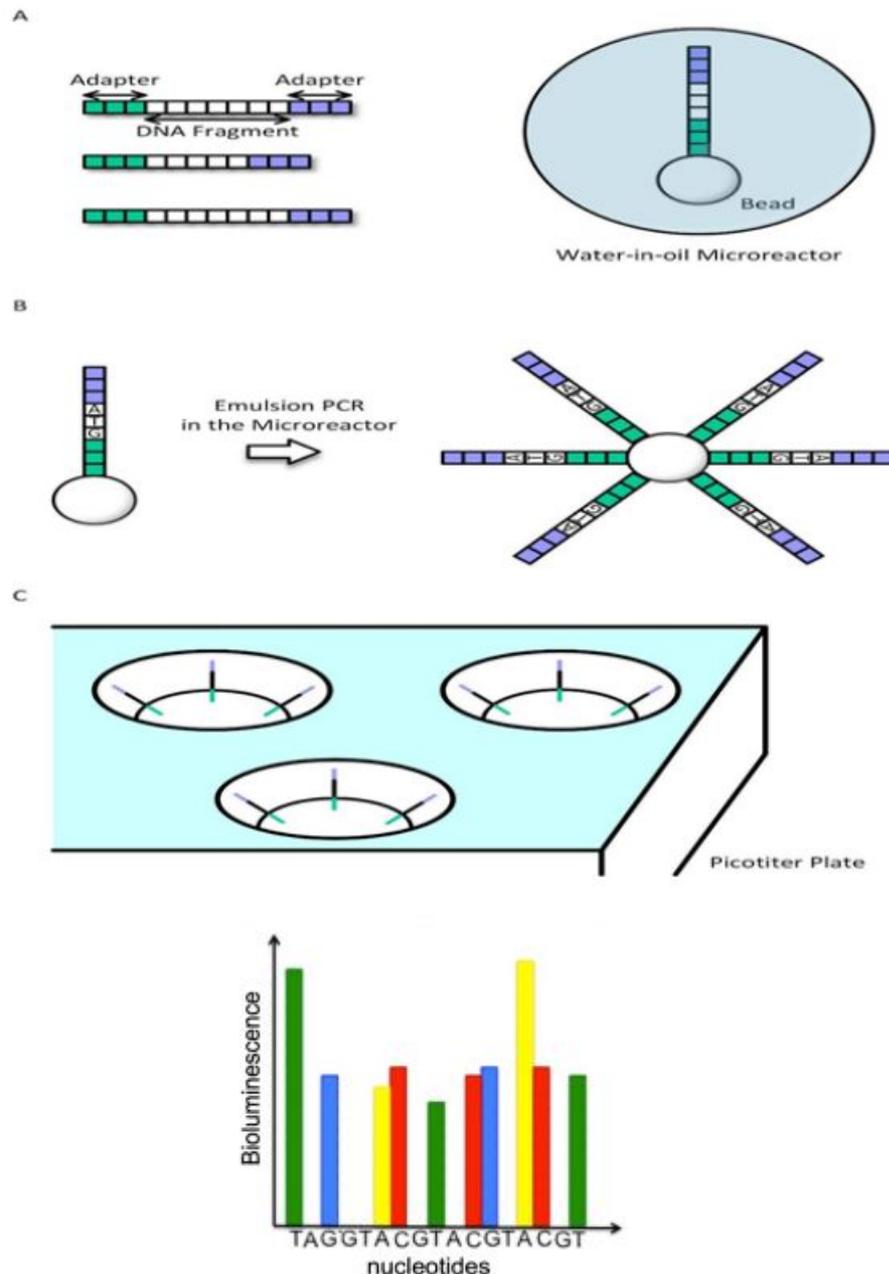
The pyrosequencing technique is a **sequencing-by-synthesis approach**.

DNA samples are randomly fragmented and each fragment is attached to a bead => each bead is associated with a single fragment.

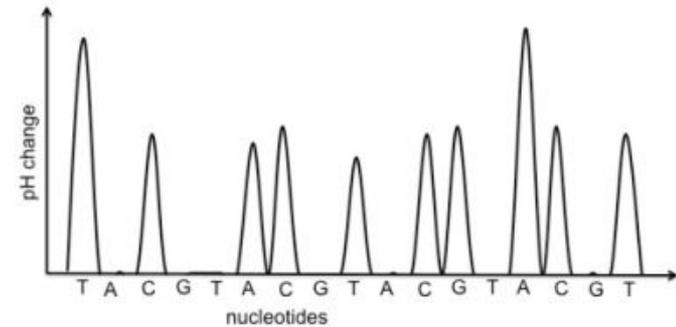
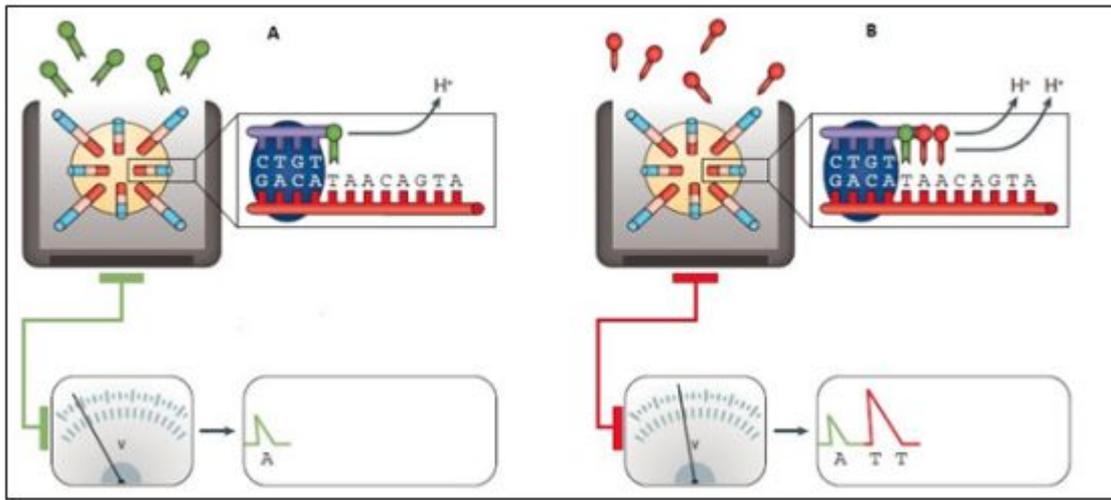
Each bead is isolated and amplified using **PCR emulsion** which produces about one million copies of each DNA fragment on the surface of the bead.

The beads are then transferred to a plate containing many wells called picotiter plate (PTP) and the **pyrosequencing technique is applied which consists in activating of a series of downstream reactions producing light at each incorporation of nucleotide. By detecting the light emission after each incorporation of nucleotide, the sequence of the DNA fragment is deduced.**

Roche 454 sequencing platform has been discontinued since 2016



Ion torrent sequencing



It is similar to 454 pyrosequencing technology but it is based on the **detection of the hydrogen ion released** during the sequencing process.

Ion Torrent uses a chip that contains a set of micro wells and each has a bead with several identical fragments. **The incorporation of each nucleotide with a fragment in the pearl, a hydrogen ion is released which change the pH of the solution. This change is detected by a sensor and converted into a voltage signal which is proportional to the number of nucleotides incorporated.**

Sequencers – Thermo Fisher Scientific

Ion Torrent next-generation sequencing systems



Ion GeneStudio S5 System

Scalable targeted NGS to support small and large projects

The Ion GeneStudio S5 system is a scalable, targeted-NGS workhorse with wide application breadth and throughput capability.



Ion Torrent Genexus System

Specimen to report in a single day with a hands-off, automated workflow*

The Genexus System is the first turnkey NGS solution that automates the specimen-to-report workflow and delivers results in a single day with just two user touchpoints.*

Illumina/Solexa

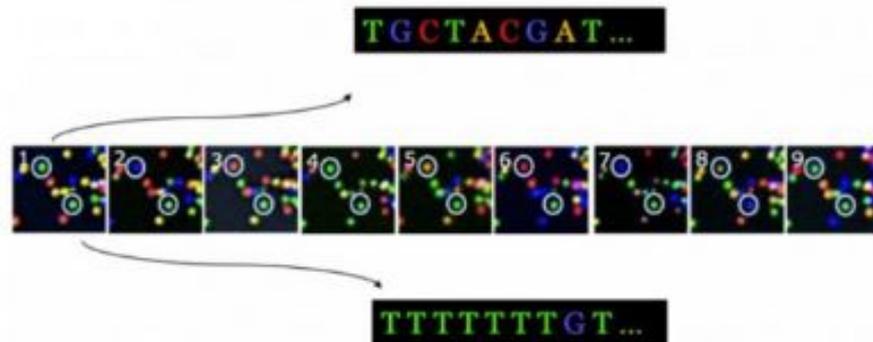
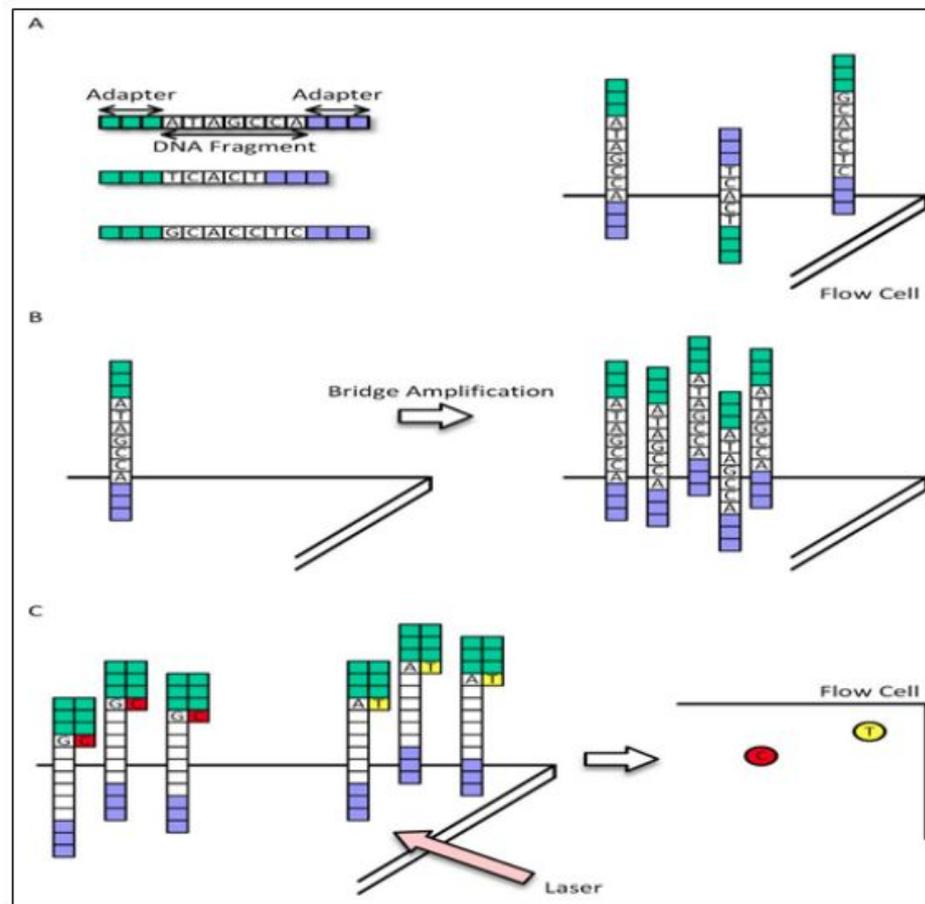
Illumina technology is **sequencing by synthesis approach** and is currently the **most used technology** in the NGS market.

The DNA samples are randomly fragmented into sequences and adapters are ligated to both ends of each sequence. These adapters are fixed themselves on a solid plate.

Each attached sequence to the solid plate is amplified by **“PCR bridge amplification”** that creates several identical copies of each sequence; a set of sequences made from the same original sequence is called a **cluster** (one million copies).

Each type of nucleotide is labeled with a fluorescent specific in order for each type to be unique. The nucleotides have an inactive 3'-hydroxyl group which ensures that only one nucleotide is incorporated. Clusters are excited by laser for emitting a **light signal specific to each nucleotide. Signals will be translated into a nucleotide sequence.**

The process continues with the elimination of the terminator with the fluorescent label and the starting of a new cycle with a new incorporation.



Benchtop sequencers



Key specifications

MiSeq i100 Series

NextSeq 550 System

NextSeq 1000 and 2000 Systems

Max output per flow cell

30 Gb^a

120 Gb^b

540 Gb

Run time (range)^c

~4–24 hr

~11–29 hr

~8–44 hr

Max reads per run (single reads)

100M^a

400M^b

1.8B

Max read length

2 × 500 bp

2 × 150 bp

2 × 300 bp

Production-scale sequencers

Key specifications



NextSeq 1000 and 2000 Systems



NovaSeq 6000 System



NovaSeq X Series

Max output per flow cell

540 Gb^a

3 Tb^b

8 Tb^c

Run time (range)^d

~8–44 hr

~13–44 hr

~17–48 hr

Max reads per run (single reads)

1.8B^a

10B (single flow cell)^b

20B (dual flow cells)

26B (single flow cell)^c

52B (dual flow cells)^{c,e}

MGI sequencers (BGI group)

Beijing Genomics Institute



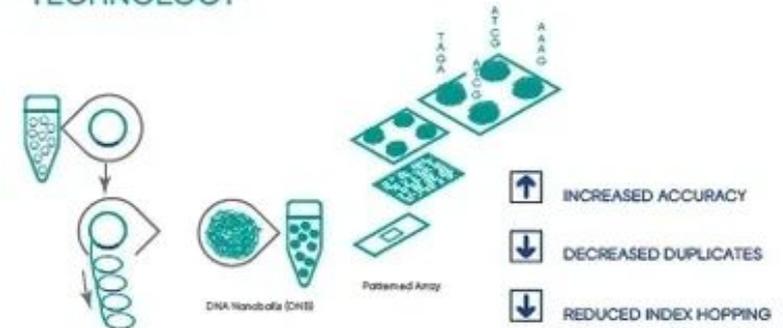
The MGISEq platform uses a unique **DNA nanoball (DNB)** technology, which involves the amplification of genomic DNA into **nanoballs, followed by sequencing by synthesis (SBS) using fluorescently labeled nucleotides.**

In nanoball sequencing, DNA fragments are amplified by rolling circle amplification. The original circular DNA fragment serves as a template for the amplification of each clonal copy of DNA. This results in a spherical "nanoball" of amplified DNA.

The **negatively charged nanoballs are then hybridized** to positively charged binding spots on an optimized patterned flow cell.

The sequencing process then proceeds in a similar fashion to standard SBS sequencing. **The nucleotides (A, C, G, or T) are added one at a time to the flow cell, and the incorporated nucleotides are detected by a camera.**

MGI'S PROPRIETARY
DNBSEQ™
TECHNOLOGY



MGI sequencers (BGI group)

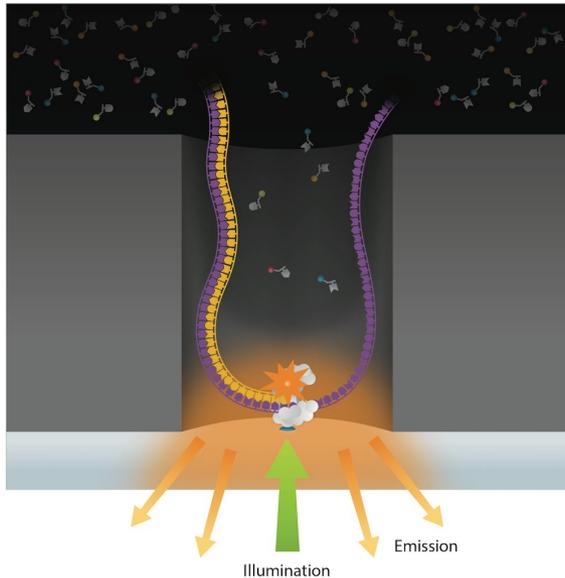


									
	Sequencers +	Sequencers +	Sequencers +	Sequencers +	Sequencers +				
Product Model	DNBSEQ-T20x2	DNBSEQ-T10x4RS	DNBSEQ-T7+	DNBSEQ-T7	DNBSEQ-T1+	DNBSEQ-G400	DNBSEQ-G50	DNBSEQ-G99	DNBSEQ-E25
Sequencing Quality	Q30>85% (PE150)	Q30>85% (PE150)	Q40>90%	Q30>85% (PE150)	Q30>93% Q40>90% (PE150)	Q30>85% Q40>85% (PE150)	Q30>80% (PE150)	Q30>90% Q40>85% (PE150, SM 2.0)	Q30>80% (PE150)
Dimensions	4200 mm *4800 mm *2000 mm	7200 mm *5000 mm *1950 mm	1,370 mm *1,760 mm *848 mm	1656 mm *903 mm *1815 mm	1150 mm*816 mm*750 mm	1086 mm *756 mm *710 mm	654 mm *489 mm *545 mm	607 mm *680 mm *640 mm	348 mm *312 mm *257 mm
Applications	Ultra-high-depth Whole Genome Sequencing	Ultra-high-depth Whole Genome Sequencing	Deep Whole Genome Sequencing, WGBS, Multi-Omics	Deep Whole Genome Sequencing	Oncology panel sequencing, WGS, WES, WGBS, etc.	WGS, WES, Transcriptome sequencing, etc.	Small whole genome sequencing, targeted panels, low-pass whole genome sequencing	Targeted oncology panel sequencing, infectious disease sequencing, 16S Metagenomic Sequencing	Rapid Identification of Pathogenic Microorganisms, Small whole genome sequencing, targeted panels
Max. Flow Cell / Run	6	8	4	4	2	2	1	2	1

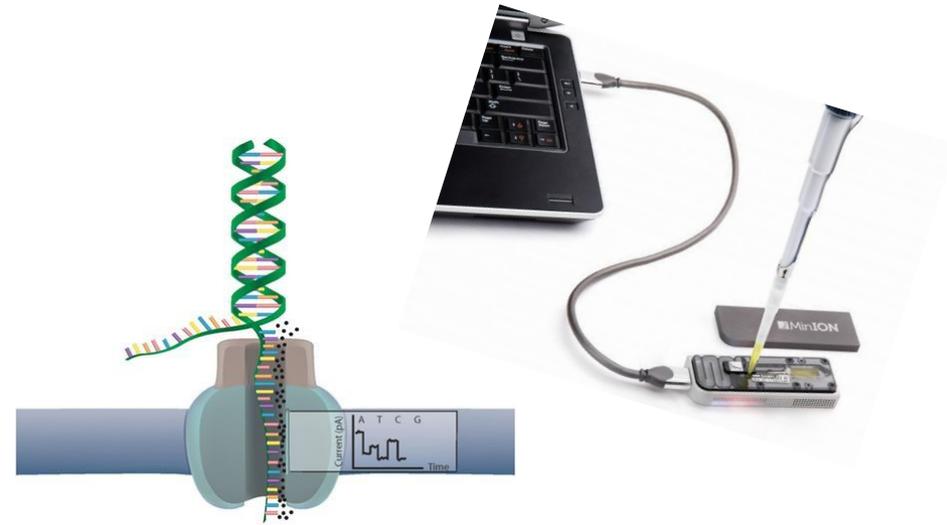
Third-generation sequencing

Third generation sequencing technologies offer the capability for **single molecule real-time sequencing of longer reads**, and detection of **DNA modification**.

PacBio SMRT technology and Oxford Nanopore can use unaltered DNA to detect methylation.



PacBio Sequencing



Nanopore technology (ONT)

- much longer reads (> Kb)
- error rate (~ 0.1 → 15 %)

Pacific biosciences SMRT sequencing

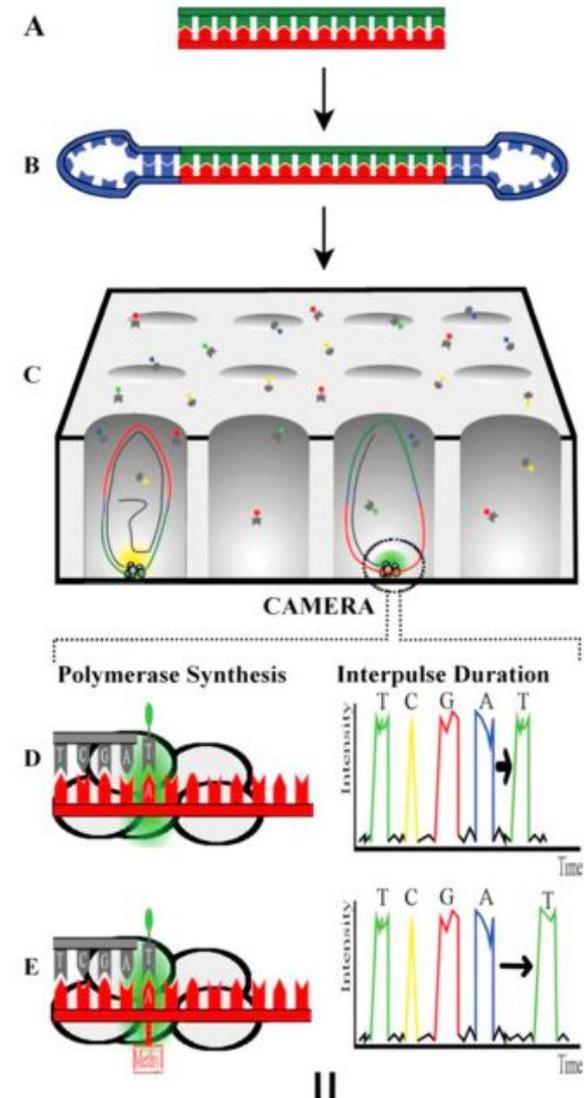
SMRT = single-molecule real-time

Sequencing starts with preparing a library from double stranded DNA (A) to which **hairpin adapters** are ligated (B).

This library is thereafter loaded onto an SMRT cell made up of nanoscale observation chambers (Zero-Mode Waveguides (ZMWs)). The DNA molecules in the library will be pulled to the bottom of the ZMW **where the polymerase will incorporate fluorescently labeled nucleotides** (C).

The fluorescence emitted by the nucleotides is recorded by a camera in real time.

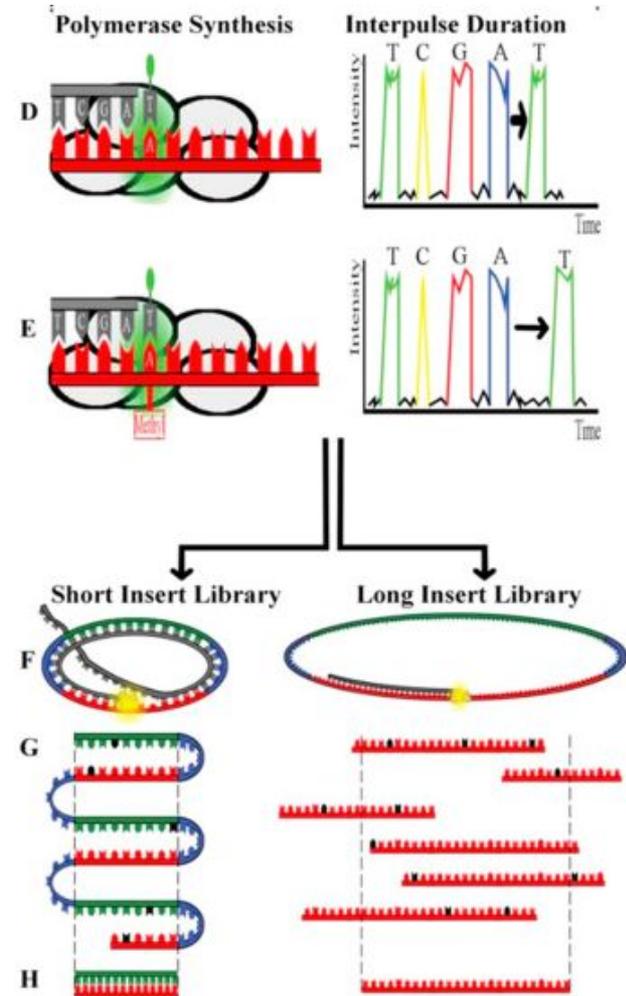
Hence, not only the fluorescence color can be registered, but also the time between nucleotide incorporation which is called the **interpulse duration** (IPD) (D, right panel). When a sequencing polymerase encounters nucleotides on the DNA strand containing an (epigenetic) modification, like for example a 6-methyl adenosine modification (E, left panel), then the IPD **will be delayed** (E, right panel) compared to nonmethylated DNA (D, right panel).



Pacific biosciences SMRT sequencing

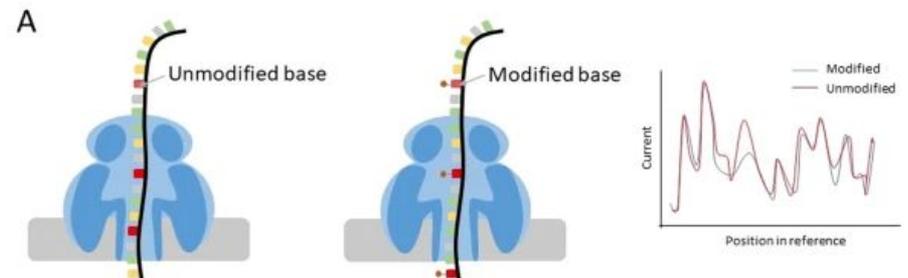
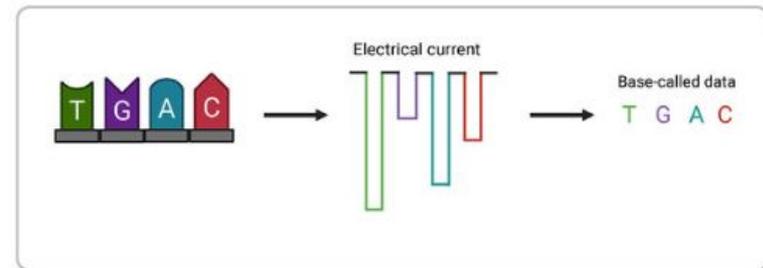
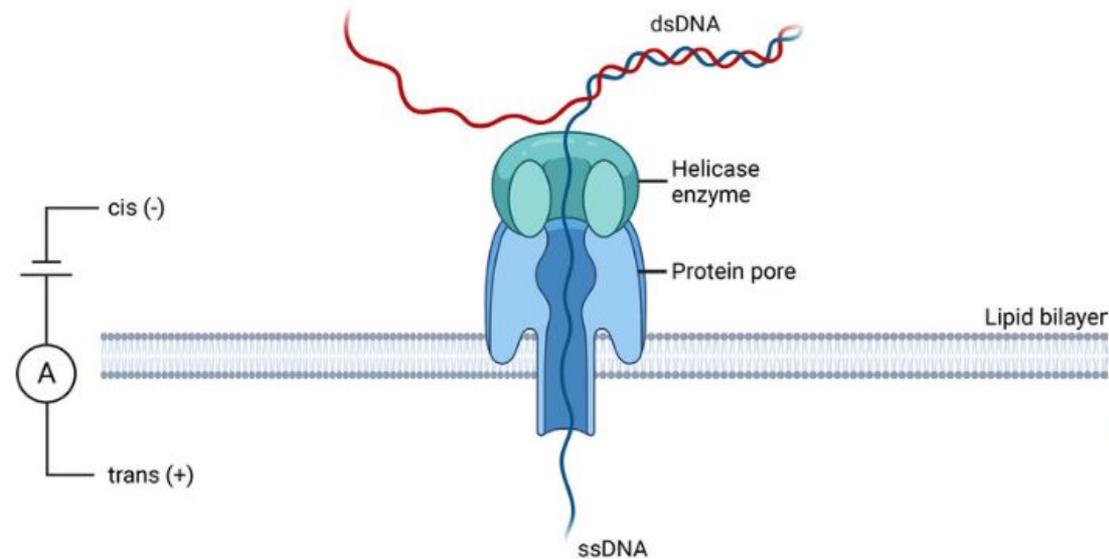
Due to the circular structure of the library, a short insert will be covered multiple times by the continuous long read (CLR). Each pass of the original DNA molecule is termed a subread, which can be combined into one highly accurate consensus sequence termed a **circular consensus sequence (CCS)** or reads-of-insert (ROI) (F–H, left panel).

Though SMRT sequencing always uses a circular template, long insert libraries typically only have a single pass and hence generate a linear sequence with single pass error rates (black nucleotides) (FG, right panel). Afterward, overlapping single passes can be combined into one consensus sequence of high quality (H, right panel). **Overall, CCS reads have the advantage of being very accurate while single passes stand out for their long read lengths (>20 kb).**



Nanopore sequencing (Oxford Nanopore Technologies)

Nanopore DNA sequencing does not require the labeling or detection of nucleotides but rather **measures the modulation of the ionic current generated when a DNA molecule passes through the nanopore.** Different nucleotides have different resistances, and measuring the time of current blockage can determine the sequence of the molecule. The technique has the potential for rapid DNA sequencing.



2019

- Different error rates and sequencing error models.
- Different output.
- Different read lengths.

Platform	Sequencing	Maximum read length (bp)	Reads per run	Run time	Maximum output	Error rate
First generation						
Sanger	NA	900	96	20 min–3 h	2.1 Mb	0.3%
Second generation						
454						
GS Junior+	Pyro	700	0.1 M	18 h	70 Mb	1% indels
GS FLX Titanium XL+	Pyro	700	1 M	23 h	700 Mb	1% indels
Illumina						
Hi Seq ^a	SBS	36 (SE)	Up to 4 B (SE)	<1–3.5 h (Hi Seq 3000/4000)	1500 Gb	0.1%
		125 (PE)	Up to 8 B (PE)	7 h – 6 d (Hi Seq 2500)		substitution
MiniSeq ^b	SBS	150 (PE)	25 M	4–24 h	7.5 Gb	<1% substitution
NextSeq 550 ^b	SBS	75 (SE)	Up to 400 M (SE)	12–30 h	120 Gb	<1%
		150 (PE)	Up to 800 M (PE)			substitution
MiSeq (v3)	SBS	75 (PE)	25 M (PE)	4–55 h	15 Gb	0.1%
		300 (PE)				substitution
Hi SeqX ^a	SBS	150 (PE)	5.3–6 B	<3 d	1800 Gb	0.1% substitution
NovaSeq6000 ^c	SBS	150 (PE)	20 B	36–44 h	6000 Gb	NA
Ion Torrent						
PGM	SBS	400 (SE)	400000–5.5 M	2.3–7.3 h	2 Gb	1% indels
Proton	SBS	Up to 200 (SE)	60–80 M	2–4 h	Up to 10 Gb	1% indels
S5	SBS	600 (SE)	2–130 M	2.5–4 h	25 Gb	1% indels

Platform	Sequencing	Maximum read length (bp)	Reads per run	Run time	Maximum output	Error rate
Oxford Nanopore						
MinION	SMRT	Up to 900 kb	Up to 1 M	Up to 48 h	20 Gb	5–10%
PacBio (Pacific Bioscience)						
RS II	SMRT	>15000 (average)	Up to 55000	30 min–4 h	1 Gb	15% indels
Sequel	SMRT	30000 (average)	~400000	30 min–20 h	10 Gb	15%

Updates

Flow cell	Kit	Sequencing & basecalling parameters	Sample	Raw read accuracy	Output
R10.4.1	Ligation Sequencing Kit V14	400 bps, 5 kHz, HAC basecalling	Human HG002	99.0% (Q20)	●●●
R10.4.1	Ligation Sequencing Kit V14	400 bps, 5 kHz, SUP basecalling	Human HG002	99.5% (Q23)	●●●
R10.4.1	Ligation Sequencing Kit V14	400 bps, 5 kHz, Duplex basecalling	Human HG002	>99.9% (Q30)	●

<https://nanoporetech.com/platform/accuracy>

Illumina launched [Complete Long Reads](#) for NovaSeq in March 2023. This kit tagments long-single-molecule fragments and can generate contiguous long-read sequences around 5-7kb in length with some reads greater than 10kb.

HOW REVIO LONG-READ SEQUENCING COMPARES

	PacBio Revio with SPRQ	SBS sequencing	Nanopore sequencing
Read length	15-20 kb ⁵	2x150 bp	10–100 kb
Read accuracy	99.95% (Q33) ¹	99.92% (Q31) ³	98.90% (Q19) ⁴
Run time	24 hours ²	44 hours	72 hours
Yield	120 - 480 Gb ²	2,400–3,000 Gb	50–110 Gb
Coverage	Unbiased	Reduced at low and high (GC)	Reduced in low-complexity runs
Genome completeness ⁶	✓	✗	✓
Variant calling – SNVs, Indels, SVs ⁵	✓	✗	✗
Areas of high homology	✓	✗	✗
Phasing	✓	✗	✓
Methylation - 5mC and 6mA	Every run	Via library prep	Every run

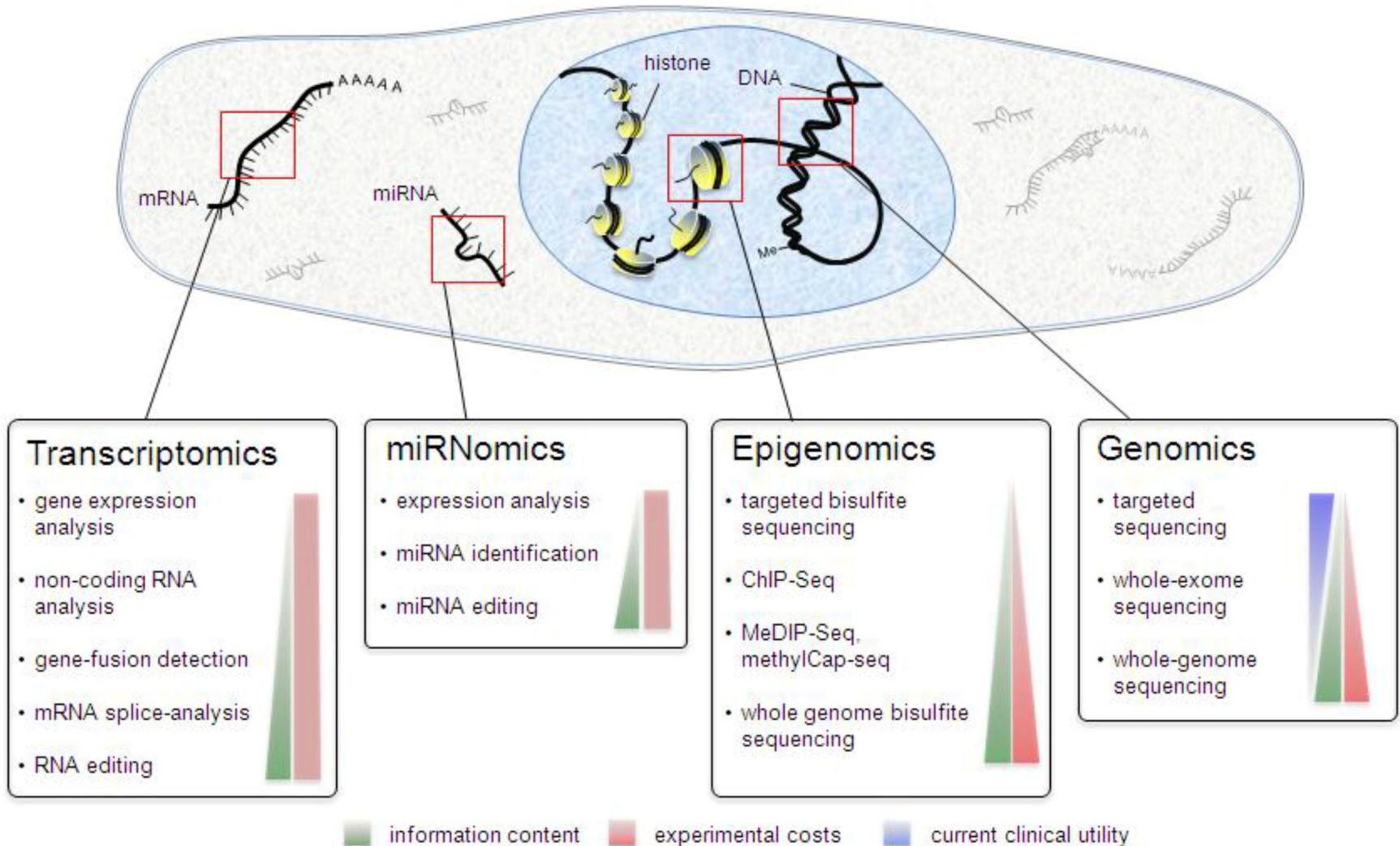
<https://www.pacb.com/revio/>

Overview of the most common sequencing platforms. WGS: whole genome sequencing; WES: whole exome sequencing; TRS: targeted sequencing, RNAseq: RNA sequencing; CCS: circular consensus sequencing; CLR: continuous long read sequencing.

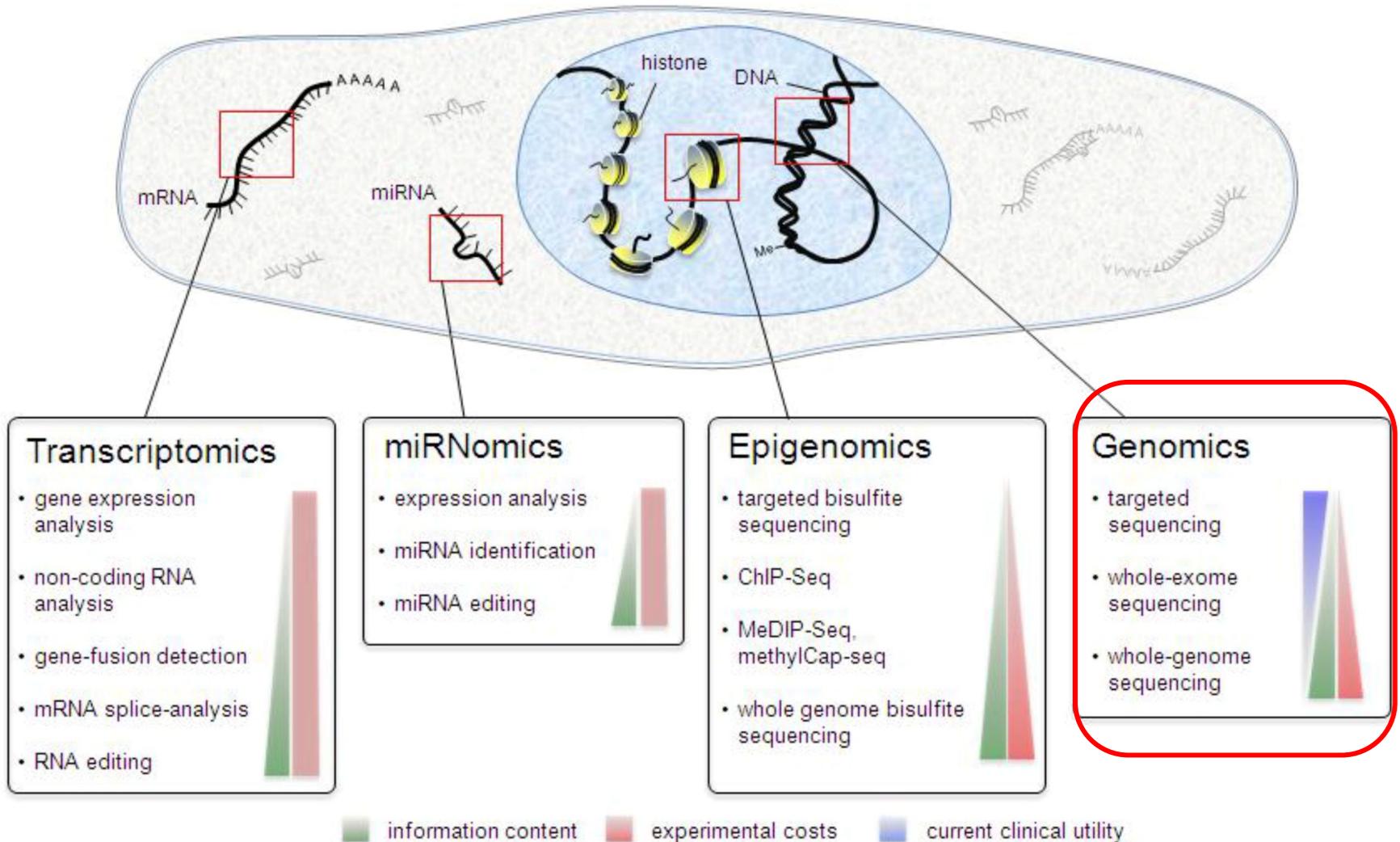
Platform	Advantages	Drawbacks	Recommended applications
Sanger	<ul style="list-style-type: none"> • Costs (low target number) • Established workflow • Simple data analysis 	<ul style="list-style-type: none"> • Sensitivity • Scalability • Sample input requirements 	TRS, validation of NGS data
Ion Torrent	<ul style="list-style-type: none"> • Costs • Speed 	<ul style="list-style-type: none"> • Short length reads • Accuracy 	TRS, metagenomics
Illumina	<ul style="list-style-type: none"> • Sensitivity • Amount of generated data with same DNA • High throughput 	<ul style="list-style-type: none"> • Costs for low target numbers • Short length reads 	WGS, WES, TRS, RNAseq, epigenomics, metagenomics
BGI Group	<ul style="list-style-type: none"> • Accuracy • No optical duplicates 	<ul style="list-style-type: none"> • Short length reads 	WGS, WES, TRS
Pacific Biosciences	<ul style="list-style-type: none"> • Long reads • High accuracy with CCS mode • Direct detection of epigenetic modifications 	<ul style="list-style-type: none"> • Costs • Large amounts of starting material • Error rate with CLR mode 	WGS, TRS, RNAseq
Oxford Nanopore Technologies	<ul style="list-style-type: none"> • Very long reads • Direct sequencing of RNA • Detection of RNA modifications 	<ul style="list-style-type: none"> • Costs • Error rate • Large amounts of starting material 	WGS*, TRS, RNAseq, epigenomics, metagenomics

*Small whole genome sequences.

Applications

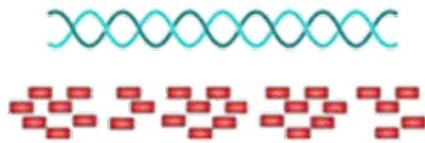


Applications



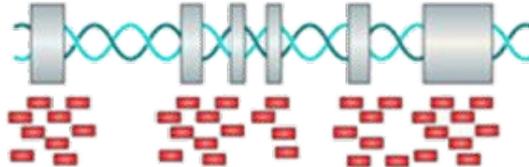
Applications : genomics (DNA-seq)

Whole genome sequencing



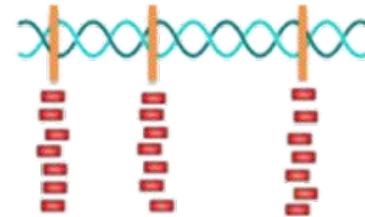
- Sequencing region : whole genome
- Sequencing Depth : >30X
- Covers everything – can identify all kinds of variants including SNPs, INDELs and SV.

Whole exome sequencing



- Sequencing region: whole exome
- Sequencing Depth : >50X ~ 100X
- Identify all kinds of variants including SNPs, INDELs and SV in coding region.
- Cost effective

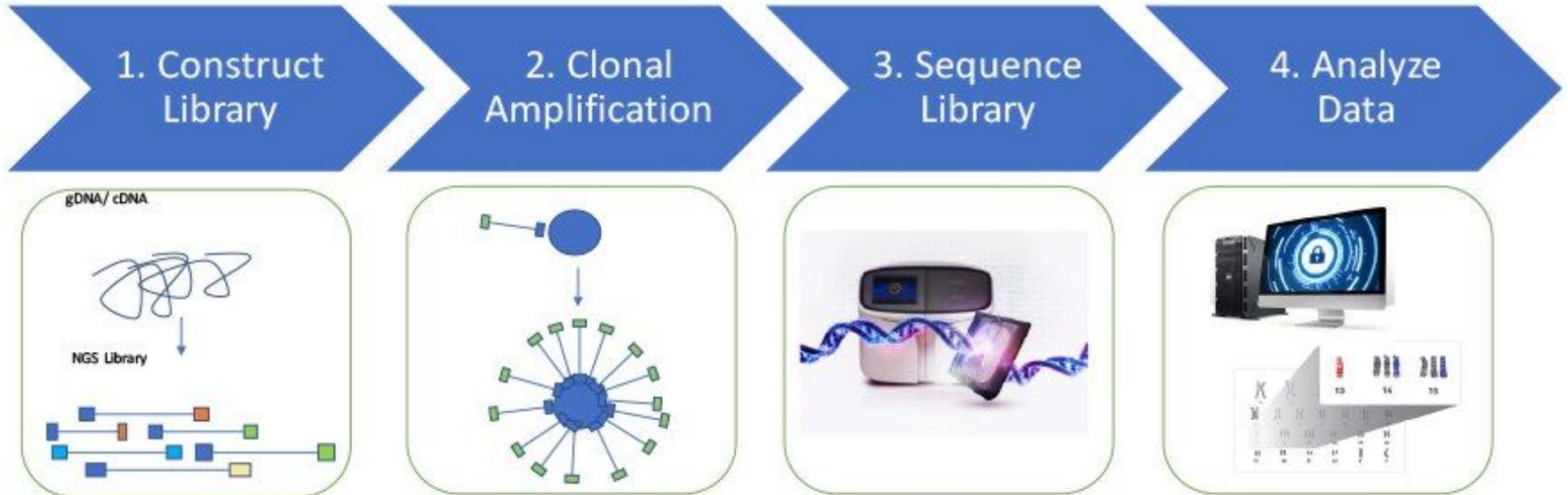
Targeted sequencing



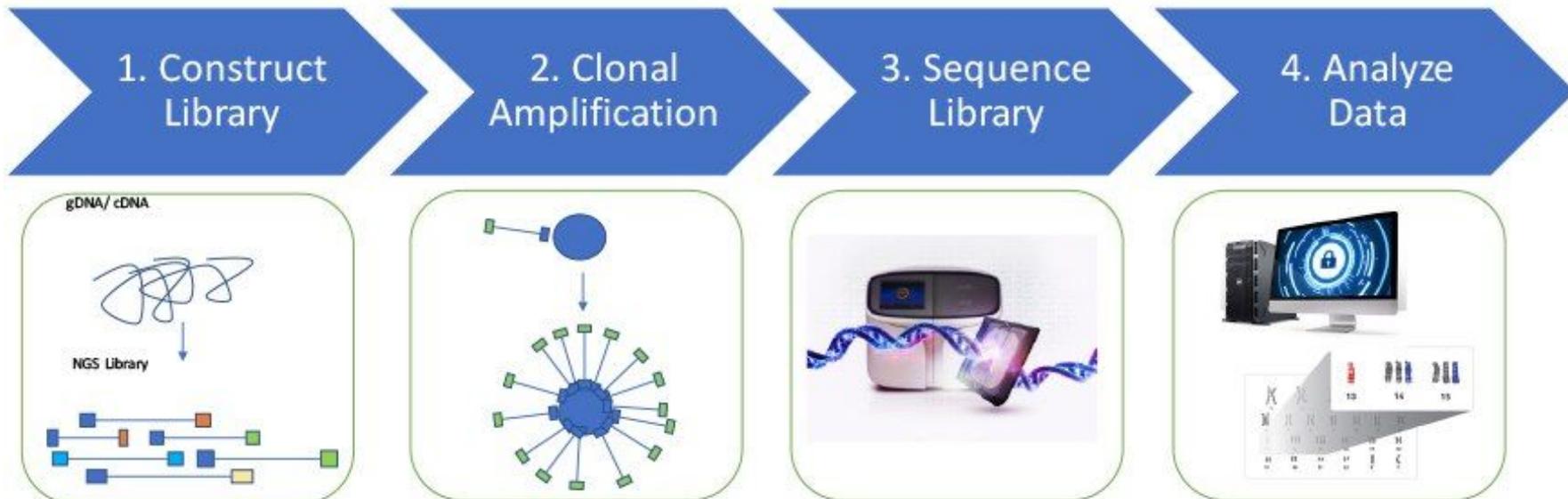
- Sequencing region: specific regions (could be customized)
- Sequencing Depth : >500X
- Identify all kinds of variants including SNPs, INDELs in specific regions
- Most Cost effective

- Targeted sequencing : rapid and cost-effective way to detect known and novel variants in selected sets of genes or genomic regions
- Whole exome sequencing : sequencing all of the protein-coding regions of genes in a genome (applications : discover rare-variants, adjacent splice-sites,...)
- Whole genome sequencing : alterations in regulatory sequences and non-coding regions, chromosomal rearrangements,

NGS workflow

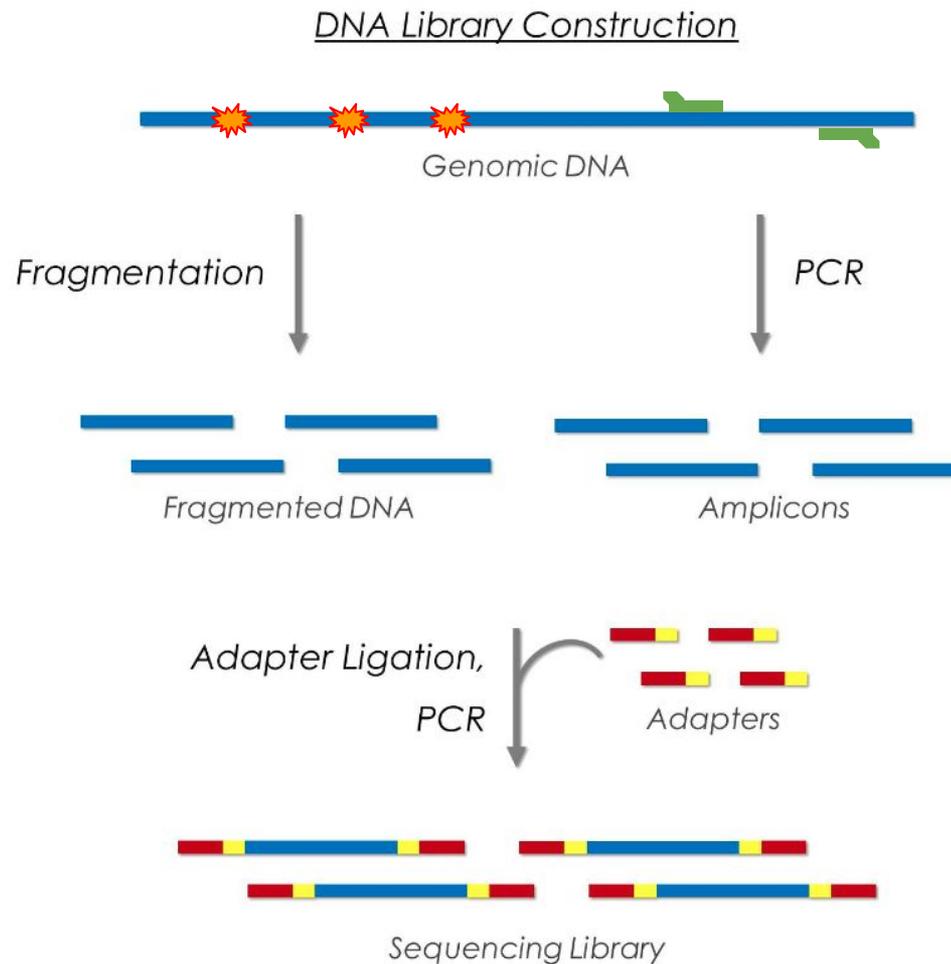


NGS workflow



Library construction

A sequencing “library” must be created from the sample. The DNA (or cDNA) sample is processed into relatively short double-stranded fragments (100–800 bp)

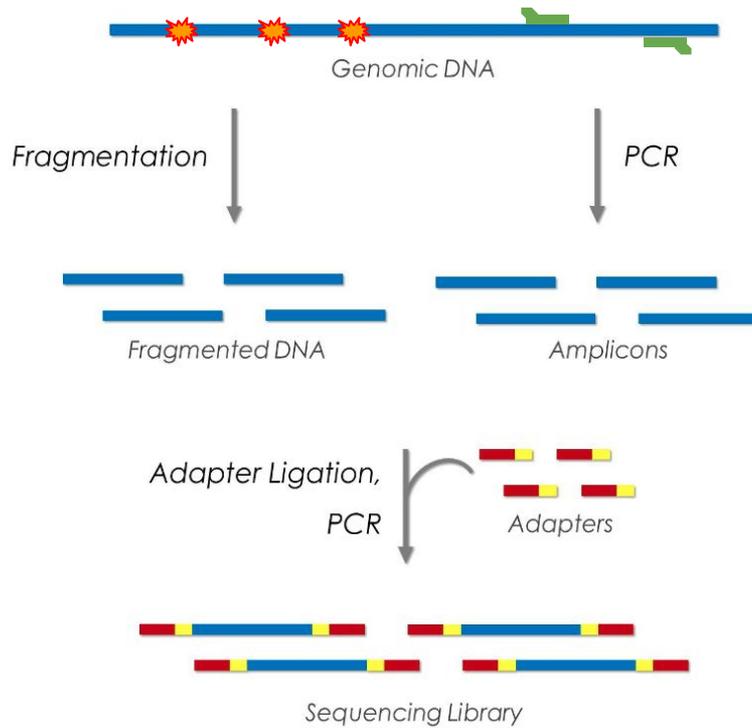


Library construction

Multiplex sequencing using DNA barcoding

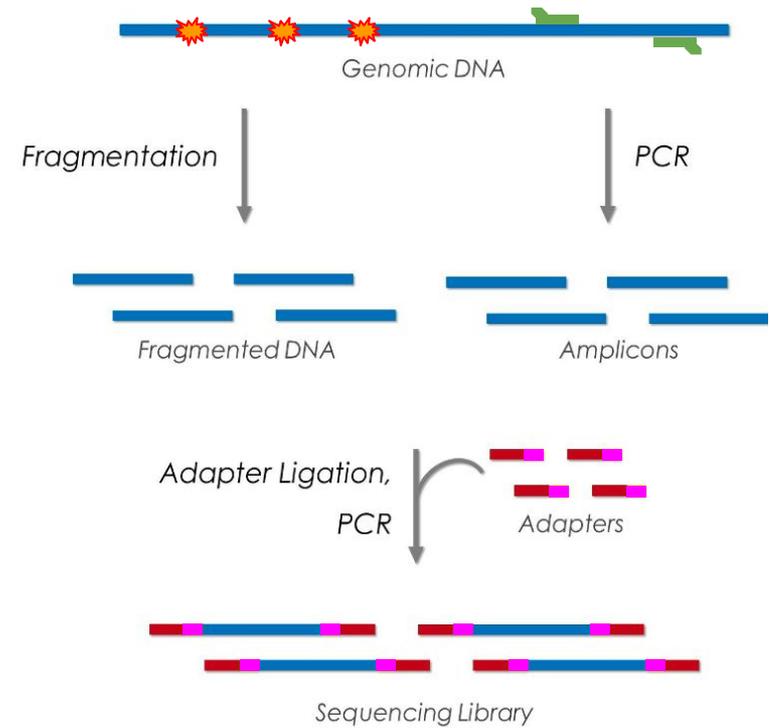
SAMPLE 1

DNA Library Construction



SAMPLE 2

DNA Library Construction



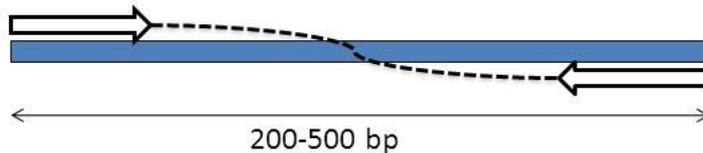
Single-end vs paired-end

- **Single-End Read:** When sequencing process only occurs in 1 direction
- **Paired-End Read:** When sequencing process occurs in both directions
- **Mate-pair Read:** Short fragments consisting of two segments that originally had a separation of several kilobases in the genome.

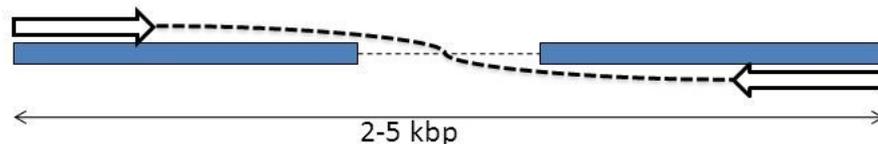
Single-End Reads - 5' or 3' (random)



Paired-End Reads - 5' and 3'



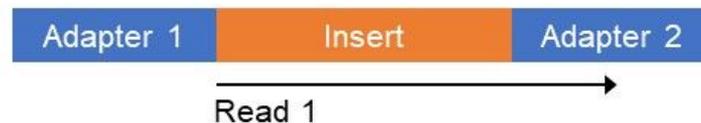
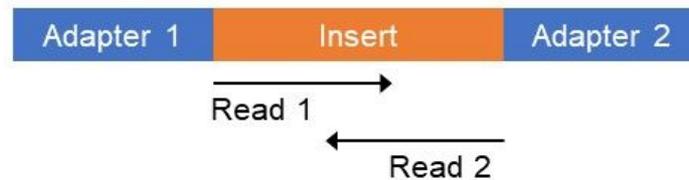
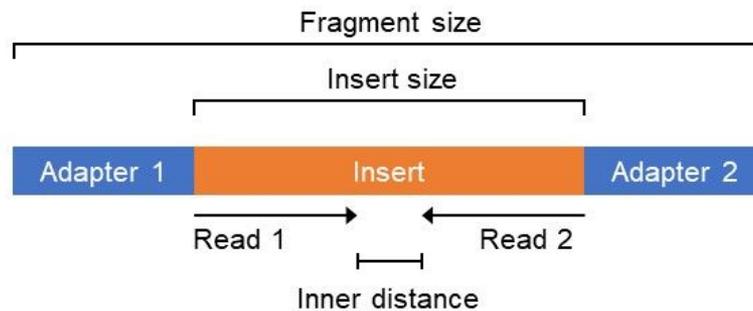
Mate-Pair Reads - 5' and 3'



source: <http://slideplayer.com/slide/7847747/25/images/7/Types+of+Sequencing+Libraries.jpg>

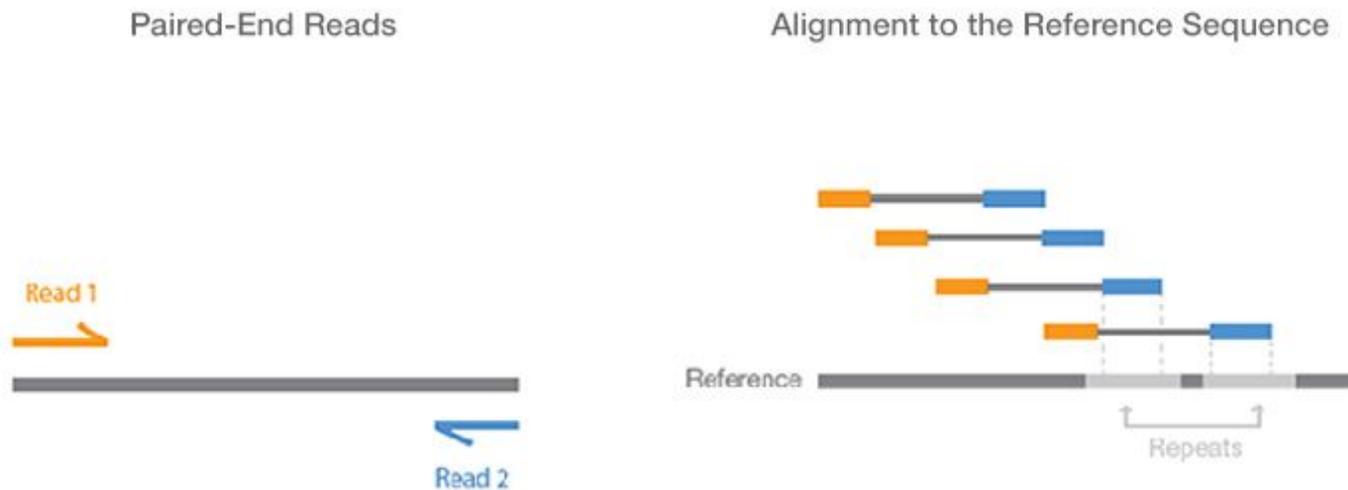
Paired-end

- **The insert size** is the size of the piece of DNA of interest, without the adapters.



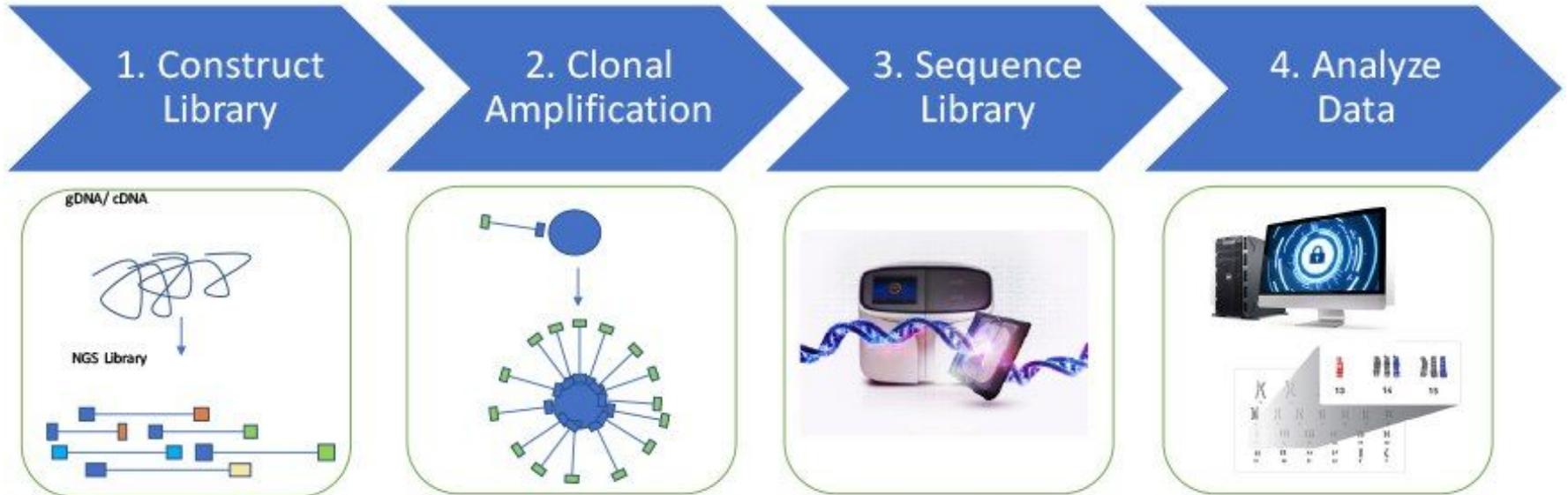
Paired-end

- **The insert size** is the size of the piece of DNA of interest, without the adapters.



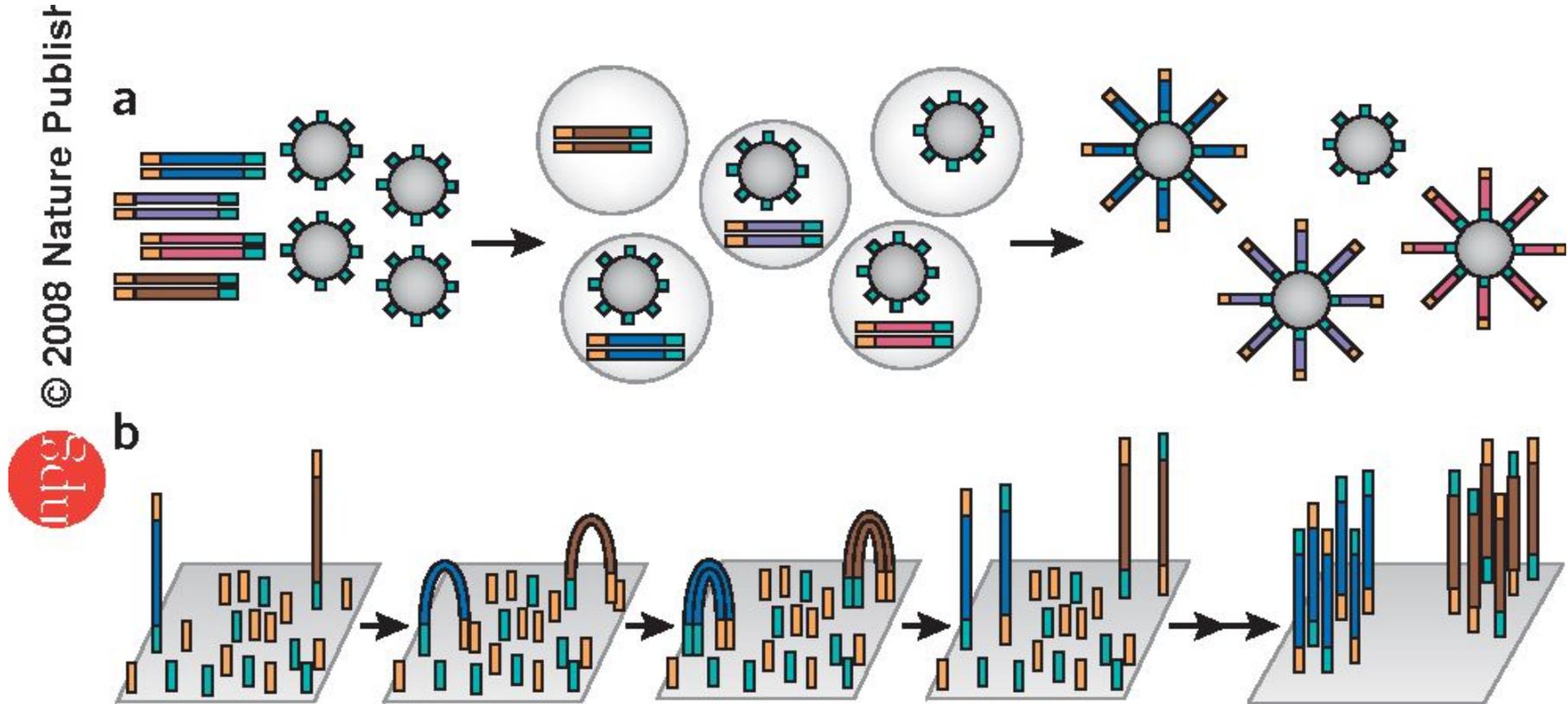
Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

NGS workflow



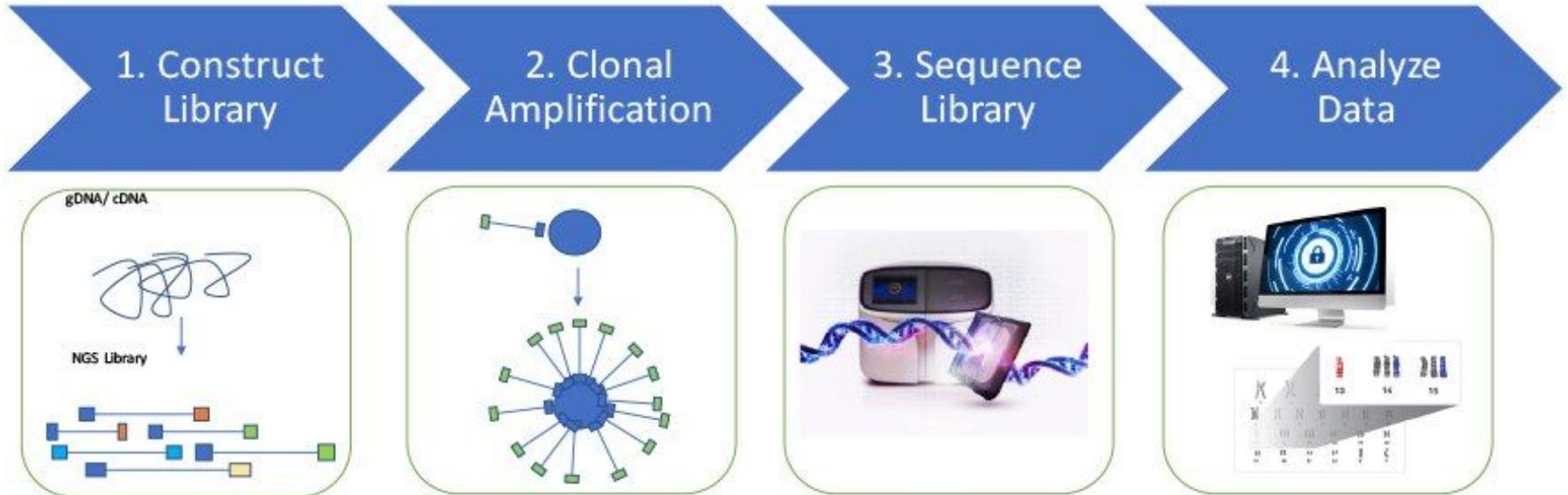
Clonal amplification

Prior to sequencing, the DNA library must be attached to a solid surface and clonally amplified to increase the signal that can be detected from each target during sequencing.



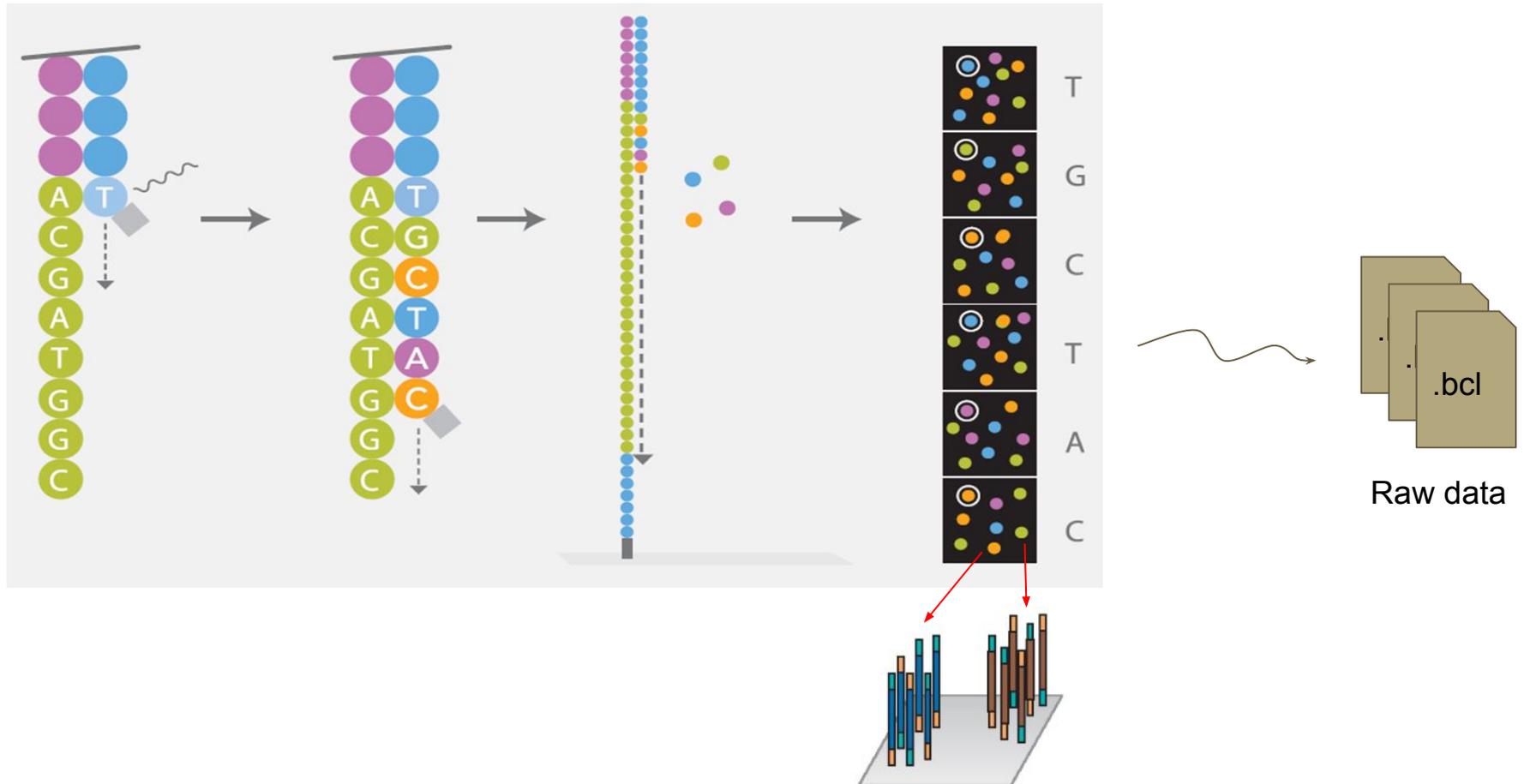
(a) thermofisher platforms rely on emulsion to amplify clonal sequencing features. (b) The Illumina technology relies on bridge PCR^{21,22} (aka 'cluster PCR') to amplify clonal sequencing features.

NGS workflow

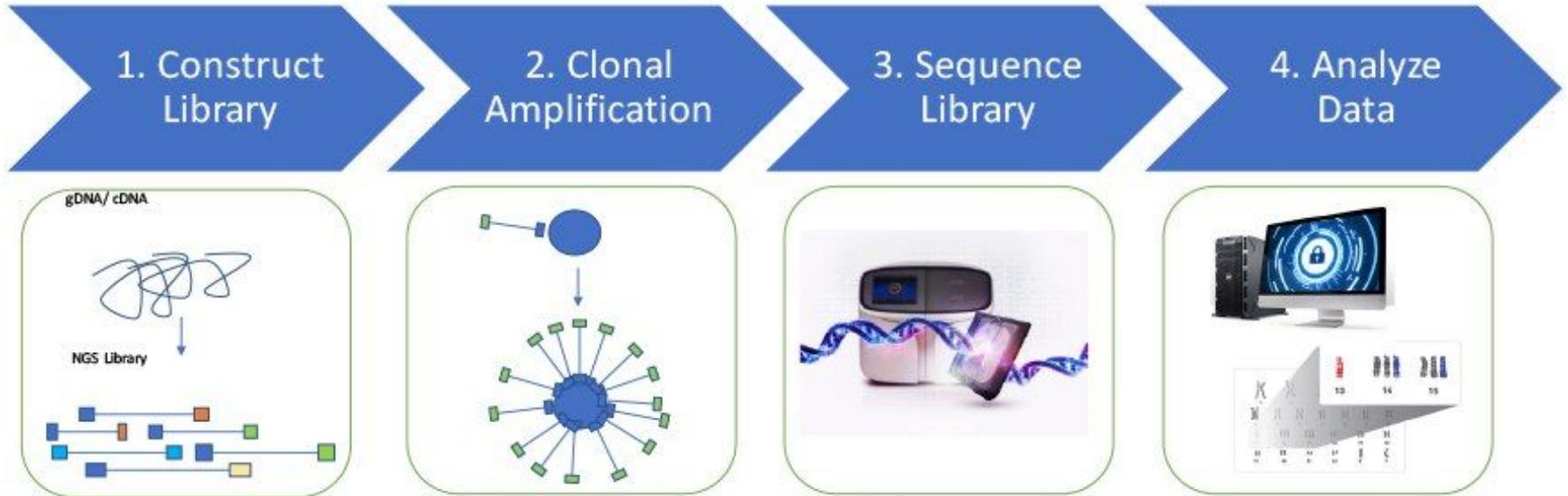


Sequencing

Illumina technology



NGS workflow

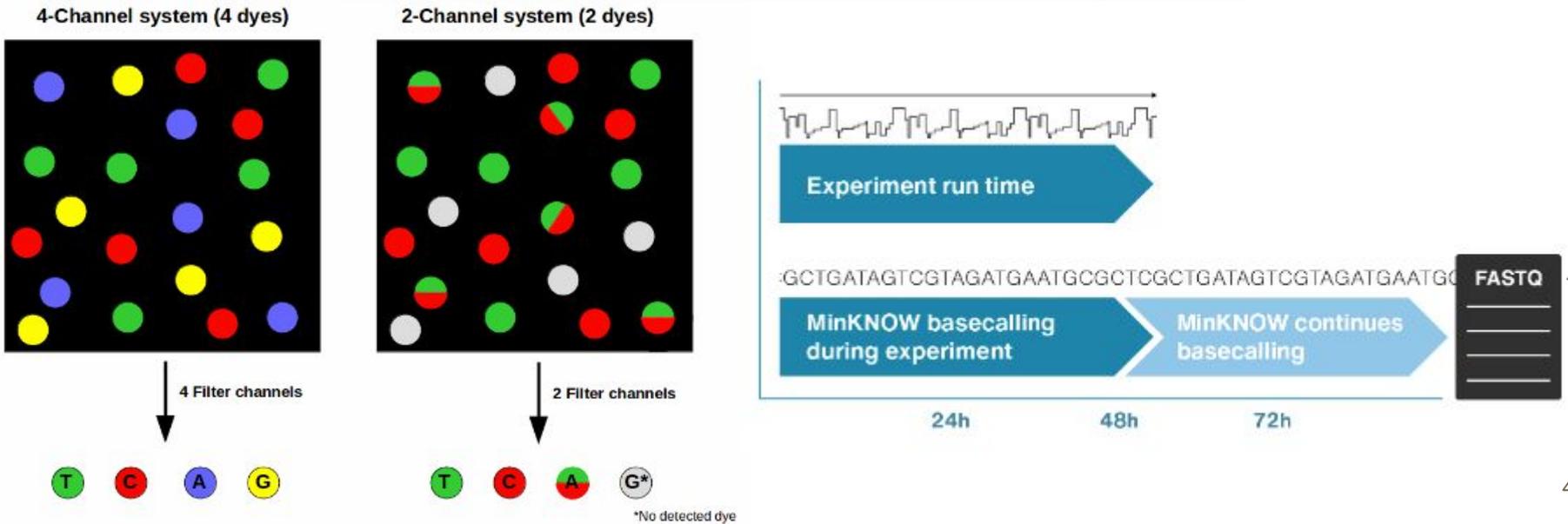


Base calling

Base calling is the process of **assigning nucleobases** to chromatogram peaks, light intensity signals, or electrical current changes.

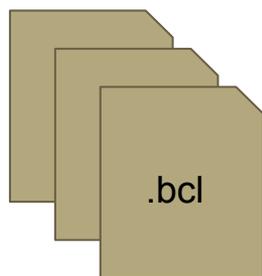


Figure: Pipeline data transformation steps (Source: Illumina)



Demultiplexing

Extracting reads, Demultiplexing



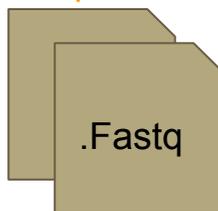
+

Sample Sheet

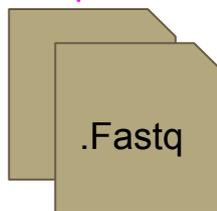
[Header]							
IEMFileVersion							
Experiment Name	Project1						
Date	4/16/2016						
Workflow	GenerateFASTQ						
Application	NextSeq FASTQ Only						
Assay	TruSeq LT						
Description							
Chemistry	Default						
[Reads]							
	151						
	151						
[Settings]							
Adapter	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA						
AdapterRead2	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT						
[Data]							
Sample_ID	Sample_Name	Sample_Plate	Sample_Well	I7_Index_ID	index	Sample_Project	Description
Sample_1				A002	CGATGT		
Sample_2				A004	TGACCA		
Sample_3				A005	ACAGTG		
Sample_4				A006	GCCAAT		

bcl2fastq

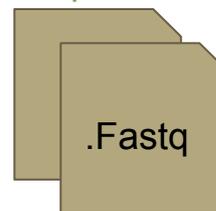
sample 1



sample 2



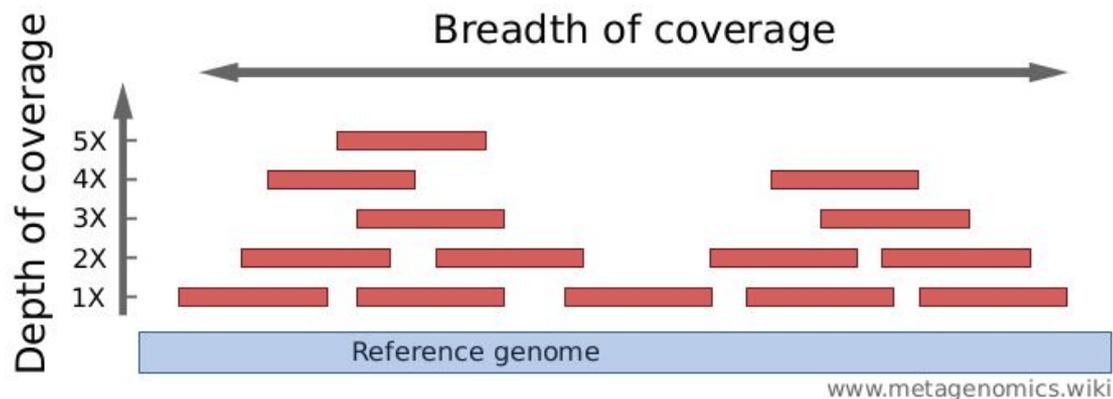
sample 3



Coverage and depth of coverage

- **Depth of coverage** = average number of reads covering a base (X)
 - Example: 30X for normal sample, 100X for tumor sample

- **(Breadth of) Coverage** = percentage of the targeted regions covered by at least X read
 - For example: 90% of a genome is covered at 1X depth; and still 40% is covered at 4X depth.



Source :

- Élodie Girard , 5ème Ecole de bioinformatique AVIESAN-IFB 2016 , http://www.france-bioinformatique.fr/sites/default/files/V01_ITMO_2016_EG_from_fastq_to_mapping_1.pdf
- <http://www.metagenomics.wiki/pdf/definition/coverage-read-depth>

Fastq file format

READ

1. Identifier

2. Sequence

4. Quality scores (as ASCII chars)

```
@SRR062641.6751359  
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT  
+  
CBLNPGJQQQJPPQPPQPQRGPPPPRRQQRPSGRQQQLRRRMEPQQPMJHQEHEKMMFIIRH?SIIHKNJIKRLJJIKHEABHIFGCGGEFCGDGDCE
```

```
@SRR062634.16249693  
CTAAGTTTGAGGCTATTTTGTTTTACAGCAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCAGAACAGGGC  
+  
ALKMOOOOPPQJQOPPPPPQPPPPPPRJRQOQQQQRPPQPRQQPFQSQQPRLLIMHKSJRQORMFELRPQNQRQJQRRPQQLIRKDMKQJPN8CFDGDCCB
```

```
@SRR062634.20060465  
CTCCCAGCTTCCAACAGACCCTGTCCCAGCTCCCTCCAAGCTGAGTGTGGCCTGATACCTACCAGTGGAGCGAGGGGAACCCGAGGACTGCCAAGGGCA  
+  
D?KMPQEPGCPQONPQIQIGR@DPERQHEKBED=HCHG8EHFD6<329@<:69A<6, ;<967>;=C:>AA8BBED#####
```

Fastq files (Paired-end)

2 files : R1, R2

Reads1.fq

```
@ERR229776.100000840
CTAGGAAGCGTAGTCTGGGGTCATCTCTCTATTAATACTGTTGGGAATGTTTAGTA
+
BAEEAGEED96EHFE@BF><>EAAC;EBH<K<6:HJGFFHBC>DDIKG4AIHFFD@0/=
@ERR229776.100020365
CATTTATTTTCATAGTAGCCAAAAAGTGAAACAGTCAAAATATCCGTCAGTGAATTGACC
+
1.*/./,/&((&3=;B@F860C>@51(3:).6GG-68C*:CG)#B4/=HDJ6;79)<@C/
@ERR229776.100104918
TATTTCTGGAATTTTCCATTTAATATTTTCAGACTGCAGTTGACTGCGGGTAACTGAAA
+
CEEEEFEDAEAGGGFDHGFHGHIIHHHIIIGKHBKJJIGHFHKILJKLEJLJJJFJMJK
```

Reads2.fq

```
@ERR229776.100000840
TTCTGGTCAGTAAGACCTCAAAGTTAAATACTAGCGATTTACACACCTTAAATGATT
+
CFIEEG@FFFGKFJHJ>HHKLLJIIJILLJIIILJHKAKJKKJJJJJLKMJKJJJKJ
@ERR229776.100020365
CCTAAAATGGTGTGTTTTTCGTATATTCACAATGCTGTGGAACCATCACCACTATCTGAT
+
4B@EDFF= (/CHBHEHCE6@ED8E@@I6HJB6E:6%@C46FFIBGCIGKD, DN=CBBE@
@ERR229776.100104918
TCTTTCTTTTGTTTTTTTTTCTGAGATGTCTTTTGTTTTTGTCTGAGGTCTTGTATG
+
CFI GGGKHHHFFHFIJIIJKLIIHJIIIKLJKKIJKLLKJFJJMHJJLJFJMJIKKJJJ
```

1 interleaved paired file

Reads.fq

```
@SRR531199.1 ILLUMINA_0130:3:1101:1249:1993 length=101
TTTTTCAGAGTAGTTGGTACCCAAATTGGAAGATGTGACCCACTTCGATACCGCGCTTGAG
+
dffffffffdfeffdadffffeeefdeffeffeffffffffffddeeYdfefefe[e
@SRR531199.1 ILLUMINA_0130:3:1101:1249:1993 length=99
ANNNNNNCTTCGGTATNAACTGGGNNNGATGTTGAACTGGGTAAAGTCGAAGATCTG
+
BBBBBBSZTUVWO]YB_[cbabbWBBBBSVVUUgggadcdedbedcddffdegeggef
@SRR531199.2 ILLUMINA_0130:3:1101:1463:1964 length=101
NTGAGTAGCTCAATGCGCTGACGCCAATAGCTATACCAACGACTGGCCAGATTATGTTT
+
BXSSRU[X[Wcc_cccccccccccc_cccccccccccccccccccccccccccc
@SRR531199.2 ILLUMINA_0130:3:1101:1463:1964 length=99
AAGTGACCATCGCGATAAAGTCTGCGCAGTAAANAGCANCTGTTNGATGCTGGCTTA
+
ggggggggggggggggggfgfgggggggggggg^BbbbaBbbaZ]BZ[ccccfgggg
@SRR531199.3 ILLUMINA_0130:3:1101:1366:1970 length=101
NAAGTCGCGGCGACCCCTATCGTGGCTTTCGGCGTACGCCATTTCAATGCGCCGCCCG
+
B[[X[YY[YVcc_cccc_cc_____][[V[^^^V[[]SXWUX[\\]]Z^^^B
@SRR531199.3 ILLUMINA_0130:3:1101:1366:1970 length=99
TGGTCAATACAAGCCGCAATACCTGCATCATGCGGNGGAANAATTTGCGCGCGTTTTCT
+
ggfegggggggdeggggfgcgggagggggggg^Bb`^]B[Y[[Zffffh_afeefe
```

Sequencing reads file formats

FastQ

READ

1. Identifier

2. Sequence

4. Quality scores (as ASCII chars)

```
@SRR062641.6751359  
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT  
+  
CBLNPGJQQQJPPQPPQPQRGPPPPRRQQRPSRGRQQQLRRRMEPQQPMJHQEHEKMMFIIRH?SIIHKNJIKRLJJIKHEABHIFGCGGEFCGDGDCE
```

```
@SRR062634.16249693  
CTAAGTTTGAGGCTATTTTGTTTTACAGCAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCCAGAACAGGGC  
+  
ALKMOOOOPPQJQOPPPPPQPPPPPPRJRQROQQQRPPRQPPQPFQSQQPRLLIMHKSJRQORMFELRPQNQRQJQRRPQQLIRKDMKQJPN8CFDGDCCCB
```

```
@SRR062634.20060465  
CTCCCAGCTTCCAACAGACCCTGTCCCAGCTCCCTCCAAGCTGAGTGTTGGCCTGATACCTACCAGTGGAGCGAGGGGAACCCGAGGACTGCCAAGGGCA  
+  
D?KMPQEPGCPQONPQIQIGR@DPERQHEKBED=HCHG8EHFD6<329@<:69A<6,;<967>;=C:>AA8BBED#####
```

FastA

```
>SRR062641.6751359  
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT
```

```
>SRR062634.16249693  
CTAAGTTTGAGGCTATTTTGTTTTACAGCAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCCAGAACAGGGC
```

```
>SRR062634.20060465  
CTCCCAGCTTCCAACAGACCCTGTCCCAGCTCCCTCCAAGCTGAGTGTTGGCCTGATACCTACCAGTGGAGCGAGGGGAACCCGAGGACTGCCAAGGGCA
```

But also: FAST5, BAM, ...

Reads Quality Control

Reads quality

- Errors when reading bases
- Depends on sequencing technologie
- Error rate tends to increase with read size

⇒ For each position in the read

- One base (A/T/C/G)
- One error probability

Phred Quality Score (for a base)

Phred quality scores Q : logarithmically related to the base-calling error probabilities P

$$Q = -10 \log_{10} P$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Quality score encoding: ASCII table

1. Identifier

2. Sequence

4. Quality scores (as ASCII chars)

READ
@SRR062641.6751359
CGCCCGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTGTTTTACAGCAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT
+
CBLNPGJQQQJPPQPPQPQRGPPPPRRRQQRPSPRRQQRLRRRMEPQQPMJHQEHEKMMFIIRH?SIIHKNJIKRLJJIKHEABHIFGCGGEFCGDGDCE

@SRR062634.16249693
CTAAGTTTGAGGCTATTTGTTTTACAGCAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCAGAACAGGGC
+
ALKMOOOOPPQJQOPPPPPQPPPPPPRJRQOQQQQRPPRQPPFQSQPRLIMHKSJRQORMFELRPQNQRQJQRRPQQLIRKDMKQJPN8CFDGGCCCB

@SRR062634.20060465
CTCCAGCTTCCAACAGACCCTGTCCAGCTCCCTCCAAGCTGAGTGTTGGCCTGATACCTACCAGTGAGCGAGGGGAACCCGAGGACTGCCAAGGGCA
+
D?KMPQEPGCPQQNPQIQIGR@DPERQHEKBED=HCHG8EHFDCD6<329@<:69A<6,;<967>;=C:>AA8BBED#####

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	\$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r

Quality score encoding

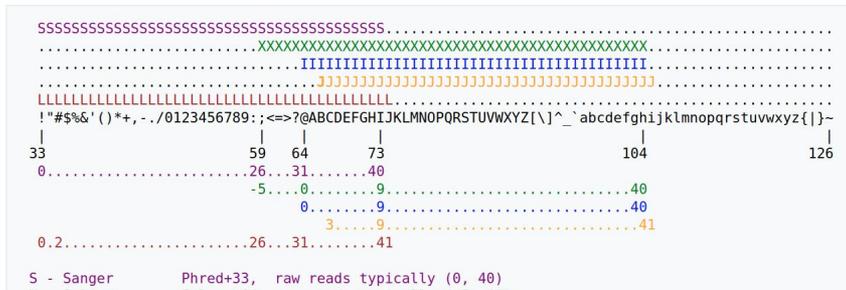
For history reasons, more than one coding convention



- S - Sanger Phred+33, raw reads typically (0, 40)
- X - Solexa Solexa+64, raw reads typically (-5, 40)
- I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
- J - Illumina 1.5+ Phred+64, raw reads typically (3, 41)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
- L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)
- P - PacBio Phred+33, HiFi reads typically (0, 93)

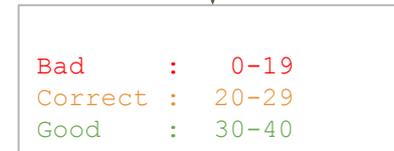
Source : <https://fr.wikipedia.org/wiki/FASTQ>

Example for score interpretation using sanger encoding



S - Sanger Phred+33

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%



Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	\$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r

```
@SEQ: ID
ACTGTACGATCGATCGCATGATCAGTACGTCGTACCAGAT
+
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|
0.....1.....2.....3.....4
01234567890123456789012345678901234567890
```

Quality Control (QC)

Quality Control (QC) is important to:

- Check if your sample sequencing went well
- Know when you need to sequence again (sequencing platform QC fail)
- Identify potential problems that can be fixed, or not
- Follow the impact of preprocessing steps

⇒ FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)

+ MultiQC (<https://multiqc.info/>) when comparing multiple datasets

FastQC Report

Summary

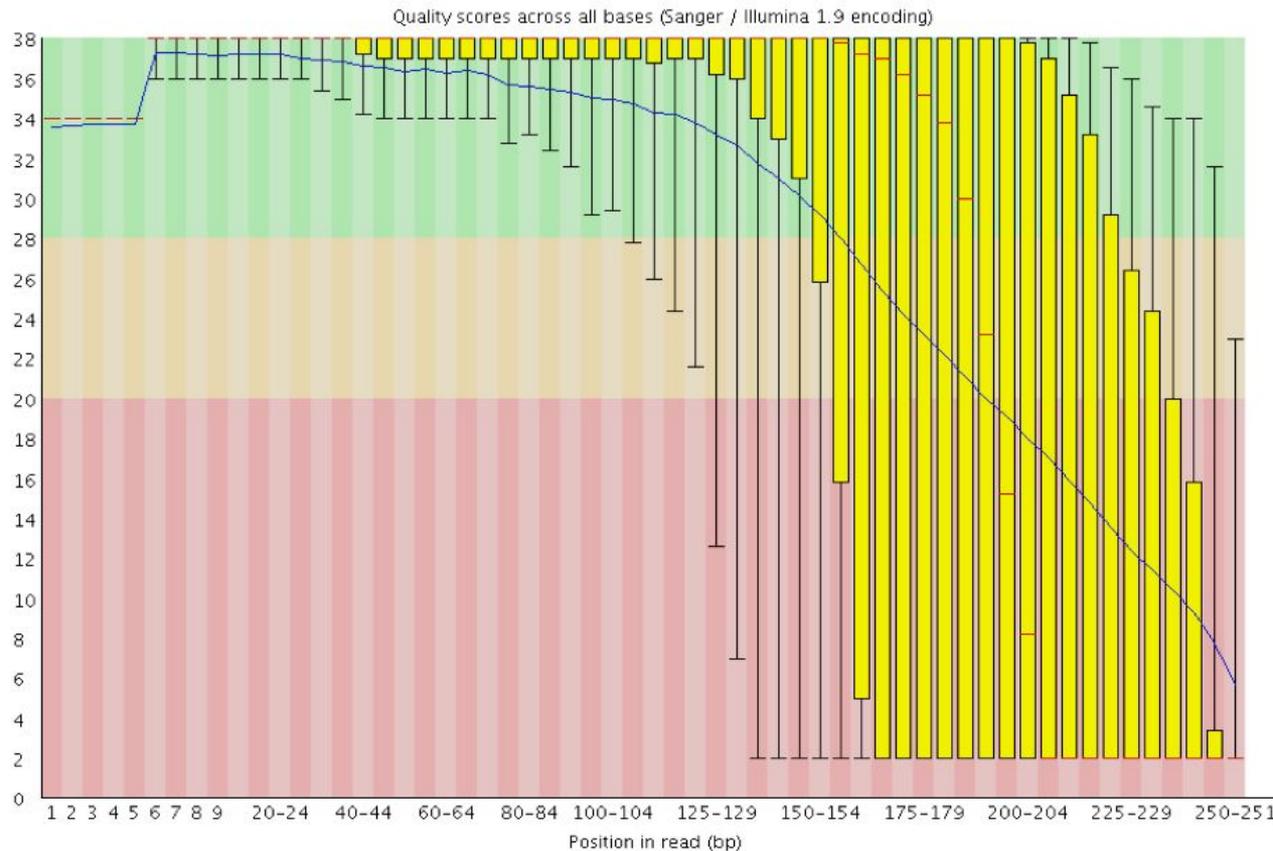
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)

Basic Statistics

Measure	Value
Filename	reads_R2_fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	45820
Total Bases	8.8 Mbp
Sequences flagged as poor quality	0
Sequence length	68-300
%GC	50

Loss of base call accuracy with increasing sequencing cycles

Source: <https://sequencing.qcfail.com>

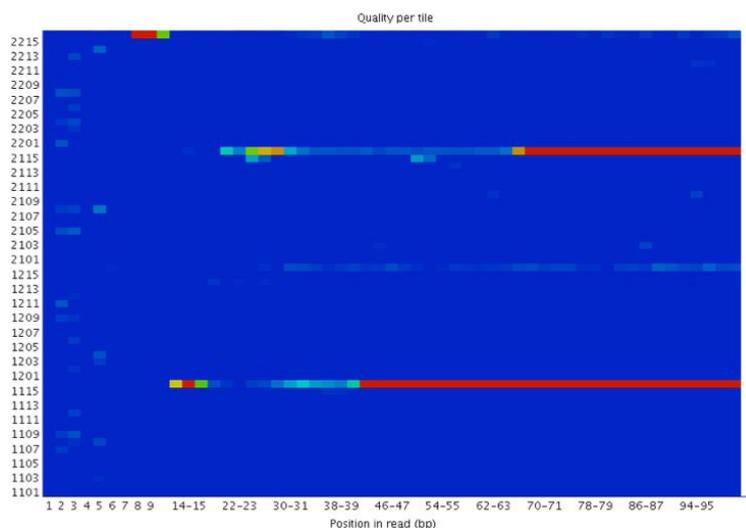
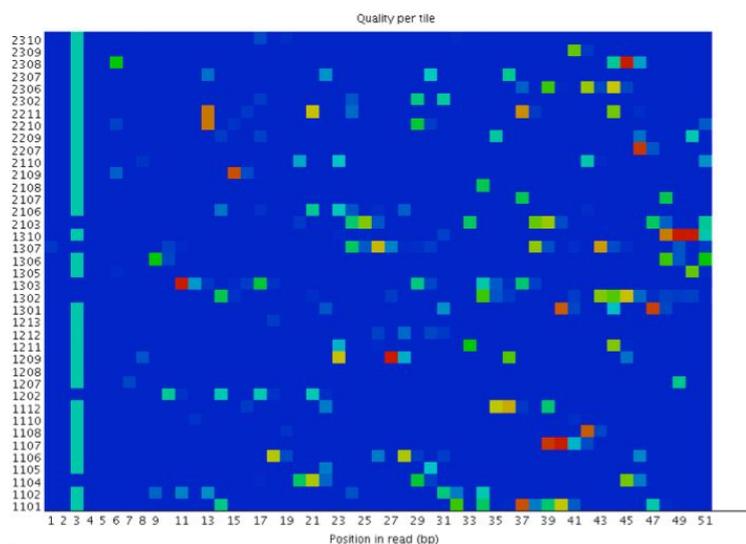


For each position, a boxplot is drawn with:

- the median value, represented by the central red line
- the inter-quartile range (25-75%), represented by the yellow box
- the 10% and 90% values in the upper and lower whiskers
- the mean quality, represented by the blue line

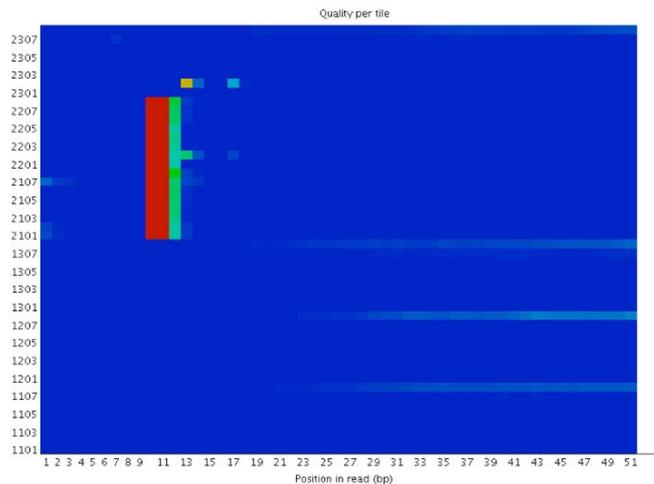
The y-axis shows the quality scores. The higher the score, the better the base call. The background of the graph divides the y-axis into very good quality scores (green), scores of reasonable quality (orange), and reads of poor quality (red).

Position specific failures of flowcells



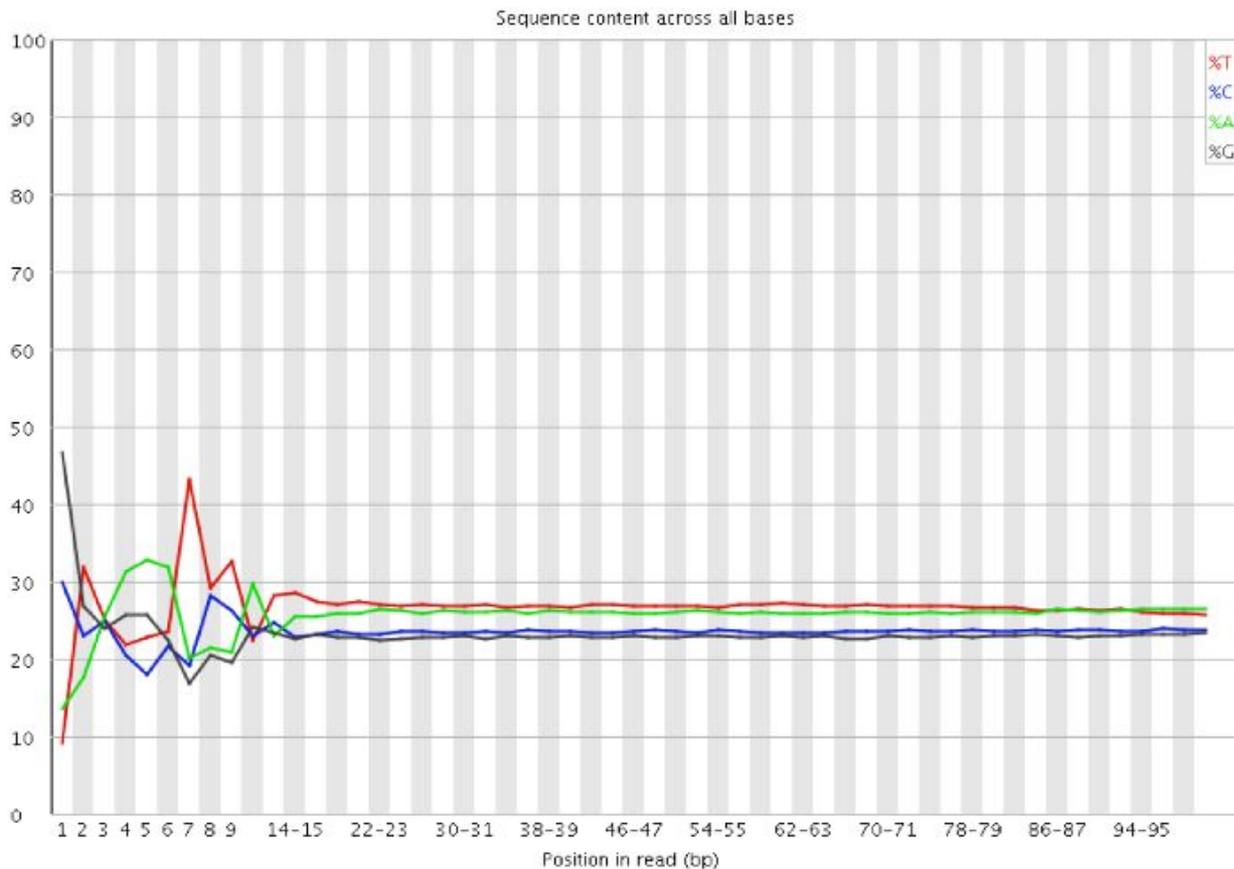
Per tile sequence quality

This plot enables you to look at the **quality scores from each tile** across all of your bases to see if there was a loss in quality associated with only one part of the flowcell. The plot shows the deviation from the average quality for each flowcell tile. **The hotter colours indicate that reads in the given tile have worse qualities** for that position than reads in other tiles. With this sample, you can see that certain tiles show consistently poor quality, especially from ~100bp onwards. **A good plot should be blue all over.**



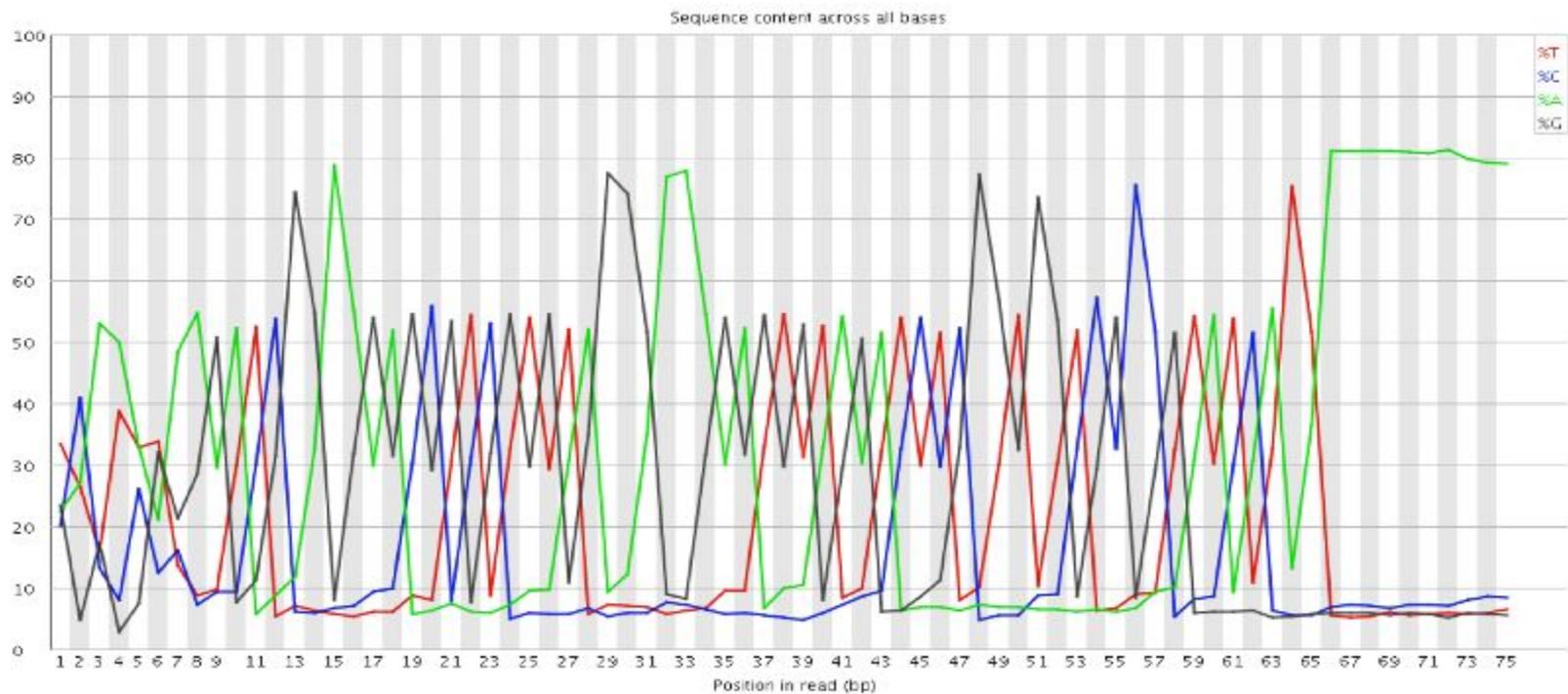
Positional sequence bias in random primed libraries

In a random library we would expect that there would be little to **no difference between the four bases**. The proportion of each of the four bases should remain relatively constant over the length of the read with %A=%T and %G=%C, and the lines in this plot should run parallel with each other.



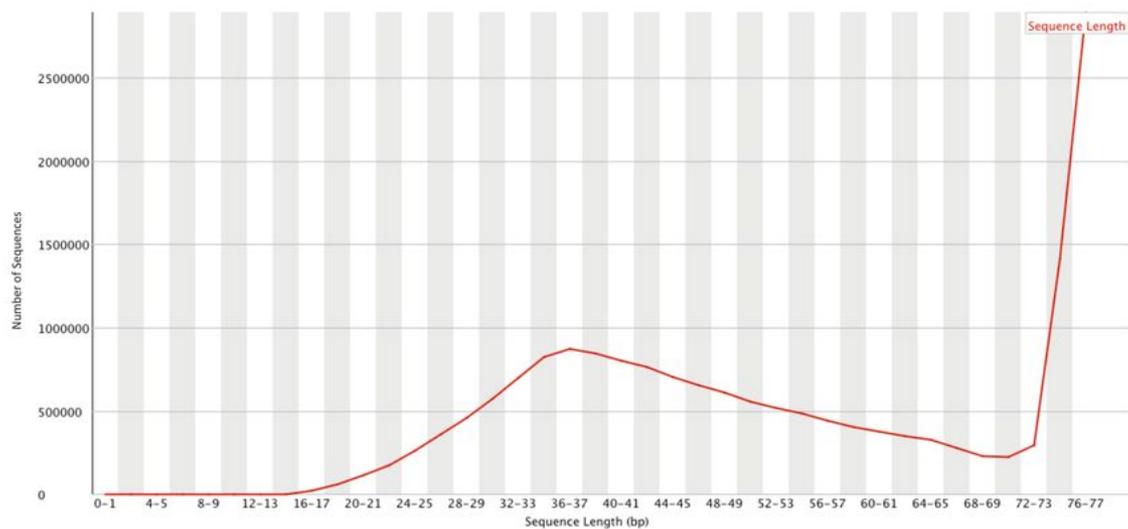
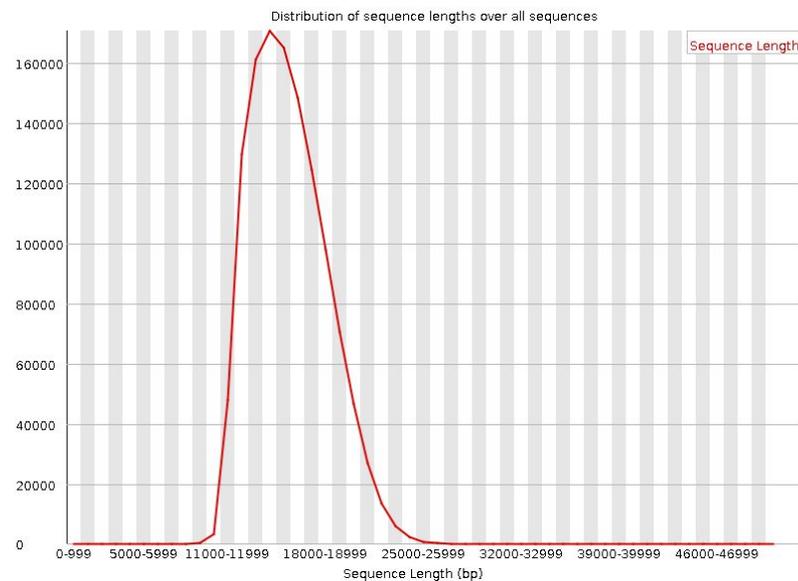
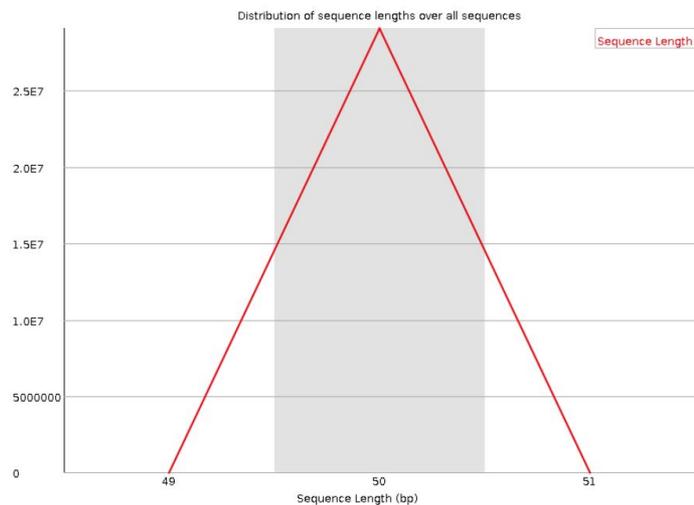
Contamination with adapter dimers

Source: <https://sequencing.qcfail.com>



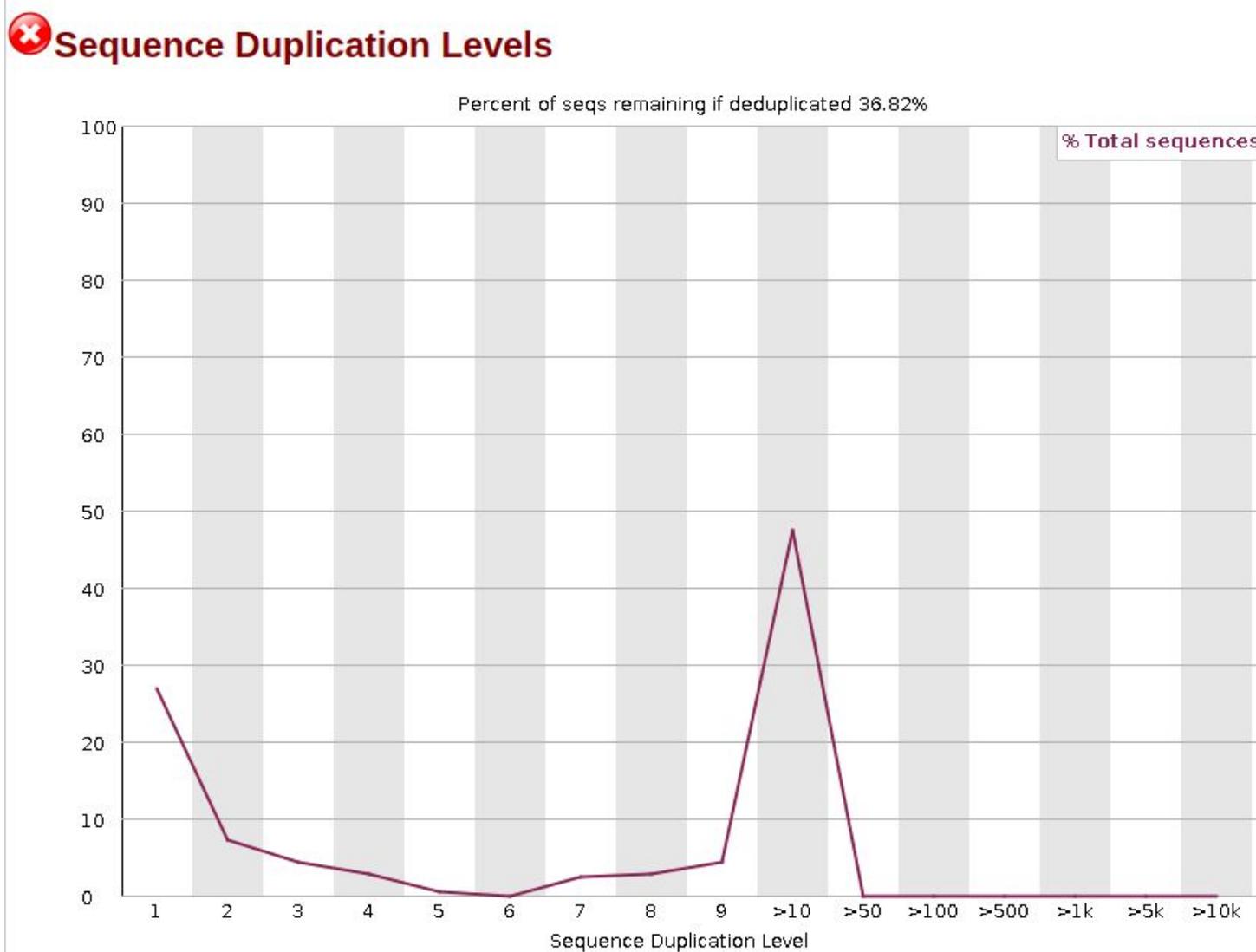
Sequence length distribution

✔ Sequence Length Distribution

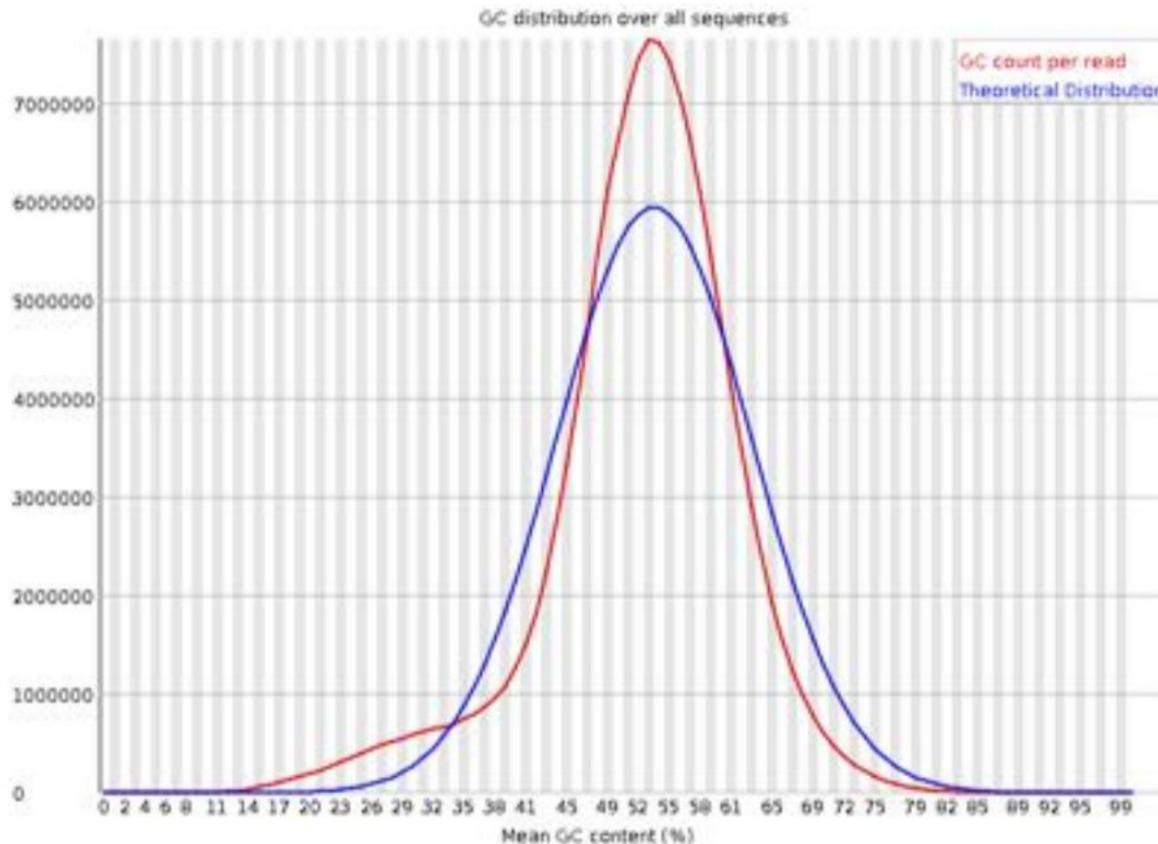


Libraries contain technical duplication

The graph shows the percentage of reads of a given sequence in the file which are present a given number of times

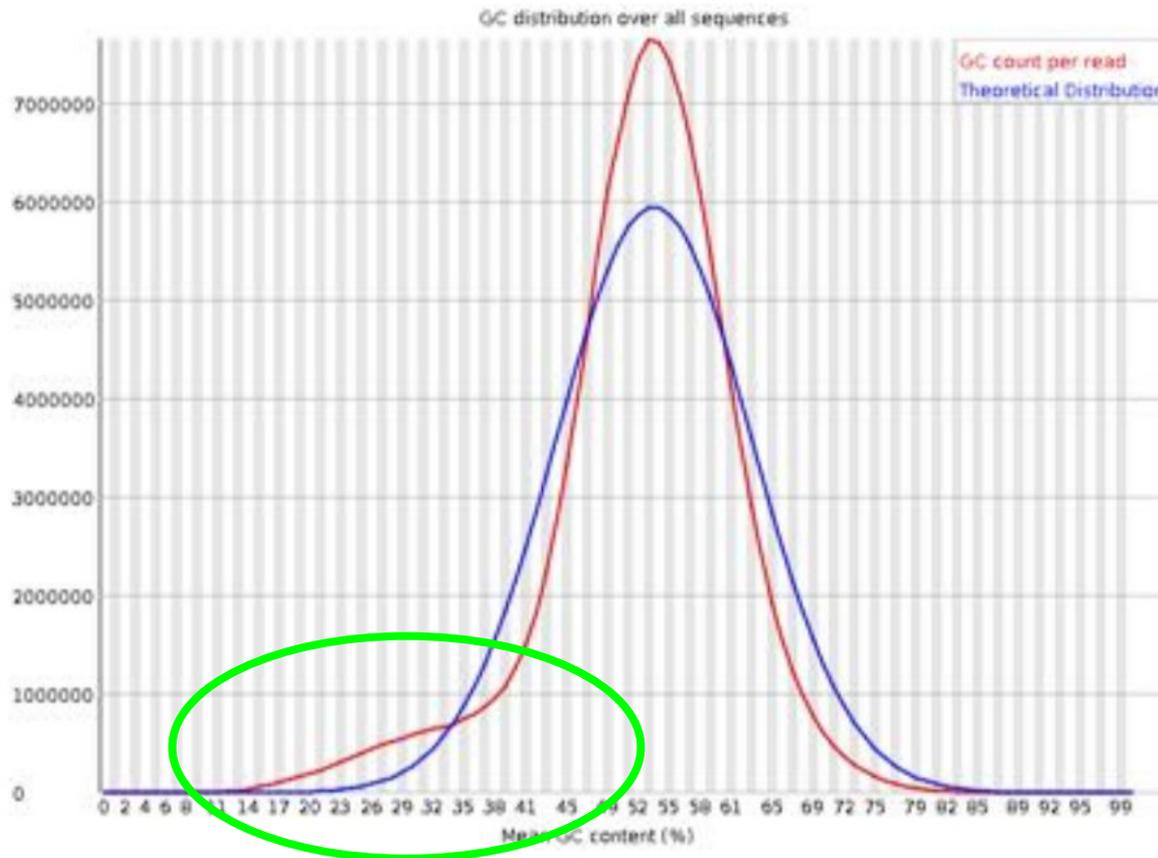


GC content / Contamination ?



This plot displays the **number of reads vs. percentage of bases G and C per read**. It is compared to a theoretical distribution assuming a **uniform GC content for all reads, expected for whole genome shotgun sequencing**, where the central peak corresponds to the overall GC content of the underlying genome. Since the GC content of the genome is not known, the modal GC content is calculated from the observed data and used to build a reference distribution.

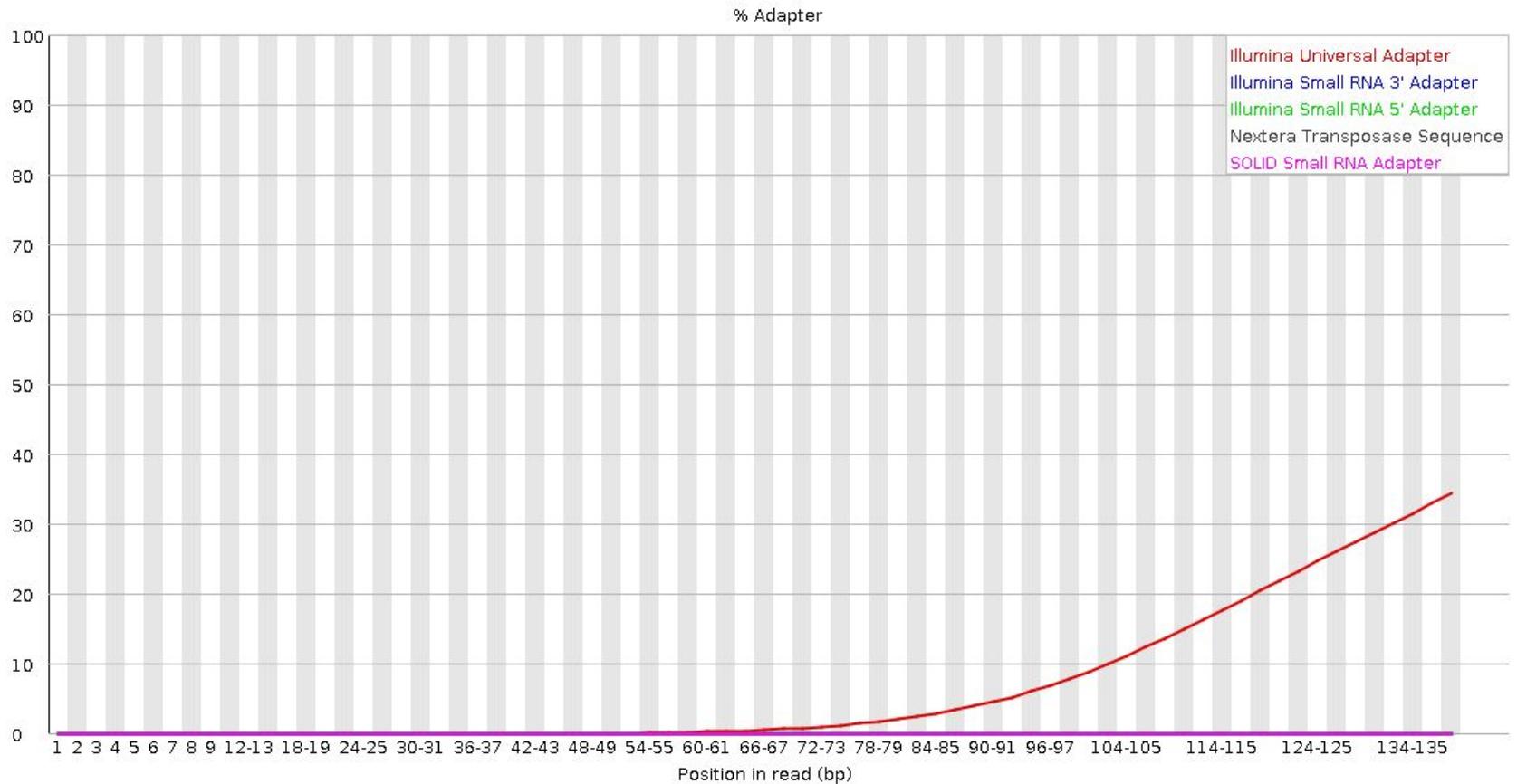
GC content / Contamination ?



This plot displays the number of reads vs. percentage of bases G and C per read. It is compared to a theoretical distribution assuming a uniform GC content for all reads, expected for whole genome shotgun sequencing, where the central peak corresponds to the overall GC content of the underlying genome. Since the GC content of the genome is not known, the modal GC content is calculated from the observed data and used to build a reference distribution.

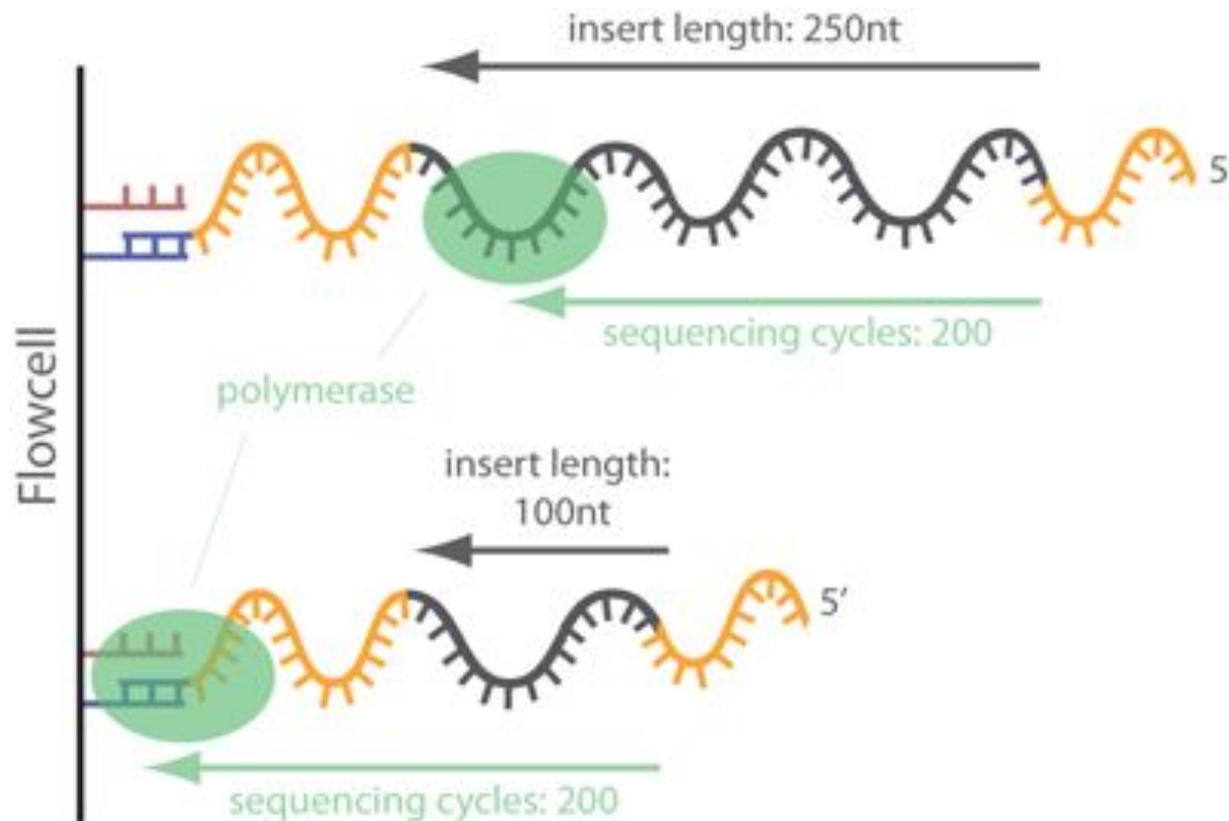
Adapter content

✖ Adapter Content



The plot shows the cumulative percentage of reads with the different adapter sequences at each position. Once an adapter sequence is seen in a read it is counted as being present right through to the end of the read so the percentage increases with the read length.

Adapter content



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGCCTTTCATCCCTTCTCAACATGAGTAAGAGAAATACGGGTAGGAAATC	6399	0.8001210372189448	No Hit
AGCCTCTCCGAGCGCGTTTCCTAAAAAGGGGGAGTCCTCATTTAAAAAAA	3452	0.43163272706357203	No Hit
ATCGGAAGAGCACACGTCTGAACTCCAGTCACTCCGCGAAATCTCGTATG	2061	0.25770424405504694	TruSeq Adapter, Index 6 (97% over 35bp)
ATGACGCTCTTCTTGAGCGTCTTTGTCTGCCGCTCTGTGCGGCTTTTT	1277	0.1596740997856841	No Hit
ATGACGCCTCTCTTTTCGGCGCTGTTTTGGAGCTTCAAAAAATGGCTGGG	1030	0.1287896028028619	No Hit
ATCGGAAGAGCACACGTCTGAACTCCAGTCACTCCGCGAAAACTCGTATG	998	0.12478837242452054	TruSeq Adapter, Index 6 (97% over 35bp)
GCCCCCTTAACATTTTCTTAACAATTTCTTAACAATCCCTACATAGTTAT	804	0.10053091325582617	No Hit

Reads Cleaning



Goal: read cleaning

```
@SRR062641.6751359
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT
+
CBLNPGJQQQJPPQPPQPQRGPPPPRRQQRPS PGRQQQLRRRMEPQQPMJHQEHEKMMFIIRH?SIIHKNJIKRLJJIKEABHIFGCGGEFCGDGDCE
@SRR062634.16249693
CTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCCAGAACAGGGC
+
ALKMOOOOPPJQOPPPPPQPPPPPRJQRQQQQRPQPRQQPFQSQQPRLIMHKS NRJQORMFELRPQNQRQJQRRPQQLIRKDMKQJRFDGCCCCB
@SRR062634.20060465
CTCCCAGCTTCCAACAGACCCTGTCCCAGCTCCCTCCAAGCTGAGTGTGGCCTGATACCTACCAGTGGAGCGAGGGGAACCCGAGGACTGCCAAGGGCA
+
D?KMPQEPGCPQQNPQIQIGR@DPERQHEKBEDHCHG8EHFDCD6<329@<:69A<6, ; <967>; =C:>AA8BBED#####
```

RAW

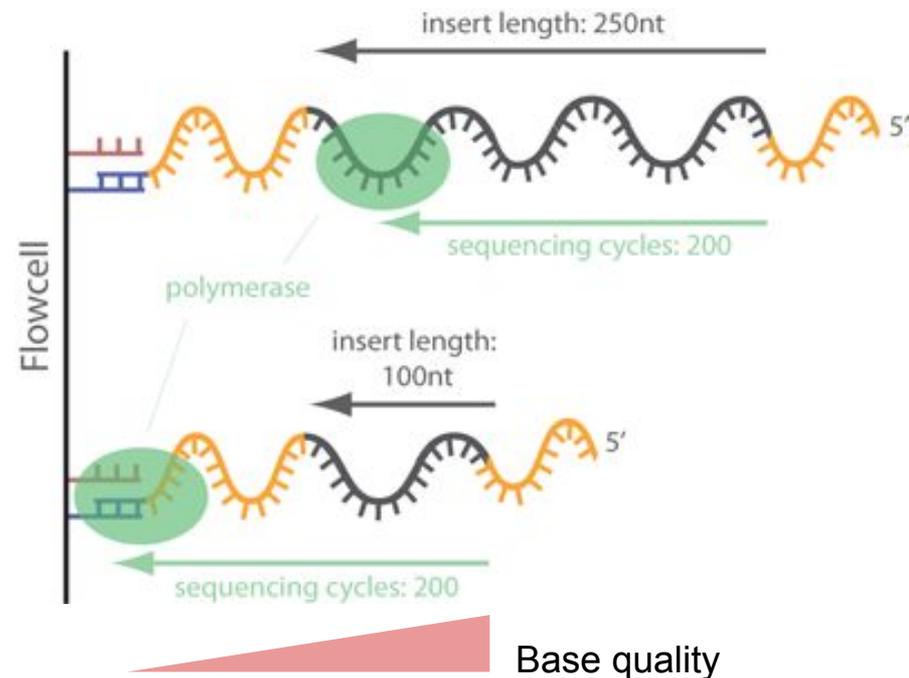


```
@SRR062641.6751359
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT
+
CBLNPGJQQQJPPQPPQPQRGPPPPRRQQRPS PGRQQQLRRRMEPQQPMJHQEHEKMMFIIRH?SIIHKNJIKRLJJIKEABHIFGCGGEFCGDGDCE
@SRR062634.16249693
CTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCCAGAACAGGGC
+
ALKMOOOOPPJQOPPPPPQPPPPPRJQRQQQQRPQPRQQPFQSQQPRLIMHKS NRJQORMFELRPQNQRQJQRRPQQLIRKDMKQJRFDGCCCCB
@SRR062634.20060465
CTCCCAGCTTCCAACAGACCCTGTCCCAGCTCCCTCCAAGCTGAG
+
D?KMPQEPGCPQQNPQIQIGR@DPERQHEKBEDHCHG8EHFDCD
```

Clean

Reads cleaning

- Cut adaptators at read ends
- Trimming : cut read ends (5' ou 3')
 - Fixed number of bases
 - Individual base quality
 - Mean quality of bases in a sliding window
- Filtering : remove read
 - Size criteria (example $< 60\text{bp}$)
 - Mean base quality for all bases criteria (example < 25)
 - Number of N



Example of quality trimming in cutadapt

We assume that the 3' end is to be quality-trimmed with a **threshold of 10** and we have the following

```
42 40 26 27 8 7 11 4 2 3
```

Step 1: Subtract the chosen threshold value from the quality value of each position

```
32 30 16 17 -2 -3 1 -6 -8 -7
```

Step 2: Compute a partial sum of these differences from the end of the sequence to each position (as long as the partial sum is negative)

```
(70) (38) 8 -8 -25 -23 -20, -21 -15 -7
```

Step 3: Cut at the minimum value of the partial sum (-25)

```
42 40 26 27
```

Reads cleaning example

Tool: Cutadapt

01

- Trimming 5' et 3' on individual base quality ($Q < 10$)

02

- Trimming of adaptaters

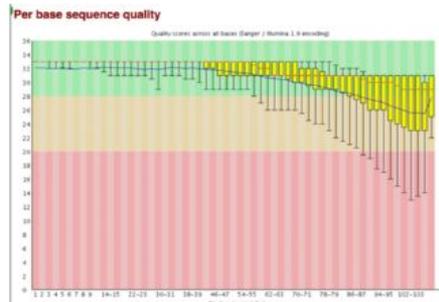
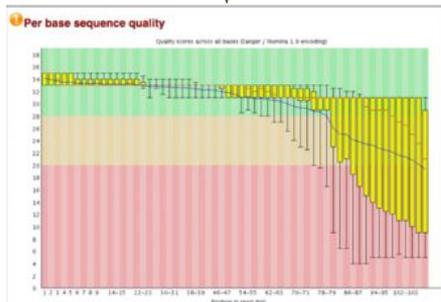
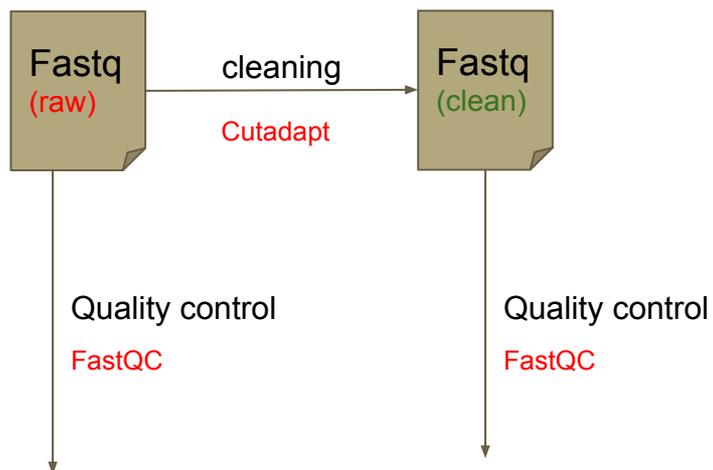
03

- Quality Trimming ($Q < 20$)

04

- Filtering on read size (taille < 50 bp)

Workflow



Practical: Quality Control (QC) & Cleaning

Open Galaxy



Practical:

<https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html>

TIAAS: <https://usegalaxy.fr/join-training/bilille-dna-2026>

N50

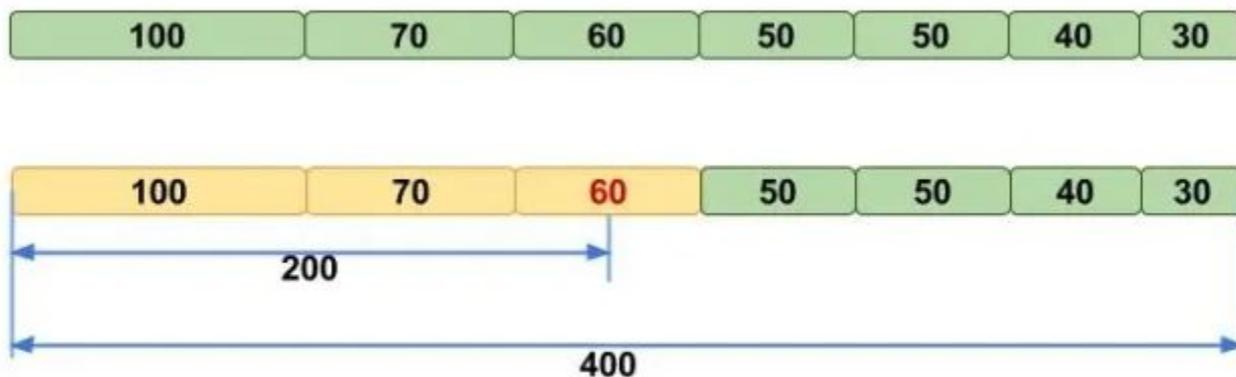


Fig. 1. Example of calculating N50 for a set of seven contigs. Here N50 equals 60 kbp.

Upper panel: Contigs, sorted according to their lengths.

Lower panel: Calculation of N50 using sorted contigs.

At least half of the nucleotides in the assembly belongs to contigs with the N50 length or longer.